# AoI-Delay Tradeoff in Mobile Edge Caching With Freshness-Aware Content Refreshing

Shan Zhang, *Member, IEEE*, Liudi Wang, *Student Member, IEEE*, Hongbin Luo, *Member, IEEE*, Xiao Ma, *Member, IEEE*, and Sheng Zhou, *Member, IEEE*

***Abstract*— Mobile edge caching can effectively reduce service delay but may introduce information staleness, calling for timely content refreshing. However, content refreshing consumes additional transmission resources and may degrade the delay performance of mobile systems. In this work, we propose a freshness-aware refreshing scheme to balance the service delay and content freshness measured by Age of Information (AoI). Specifically, the cached content items will be refreshed to the up-to-date version upon user requests if the AoI exceeds a certain threshold (named as refreshing window). The average AoI and service delay are derived in closed forms approximately, which reveals an AoI-delay tradeoff relationship with respect to the refreshing window. In addition, the refreshing window is optimized to minimize the average delay while meeting the AoI requirements, and the results indicate to set a smaller refreshing window for the popular content items. Extensive simulations are conducted on the OMNeT++ platform to validate the analytical results. The results indicate that the proposed scheme can restrain frequent refreshing as the request arrival rate increases, whereby the average delay can be reduced by around 80% while maintaining the AoI below one second in heavily-loaded scenarios.**

***Index Terms*— Mobile edge caching, age of information (AoI), delay, cache refreshing, content dynamics.**

## I. Introduction

MOBILE edge caching enables content service in proximity by using the storage resources of radio access facilities and mobile devices, bringing tremendous benefits for both network users and operators [2]. The in-proximity service can effectively reduce the end-to-end delay and better support time-critical applications [3]–[6]. Besides, the backhaul transmission pressure is relieved with reduced duplicated transmissions, whereby the capacity can be effectively enhanced especially in the backhaul-constrained dense networks [7]. Furthermore, assisted with big data analysis, the contents can be cached pro-actively based on user preference, mobility and behaviors, providing fine-grained customized services [8], [9]. With these attractive potential benefits, mobile edge caching is considered as a cornerstone of the 5G and beyond networks [10], [11].

Despite the attractive benefits, mobile edge caching may lead to staleness of dynamic items whose content information changes with time and environment [12]. Examples can be the content of the same URL, real-time street maps, traffic congestion of a certain region, temperature and air conditions of a room, etc. Therefore, the cached items should be refreshed timely to the most recent versions depending on the content dynamics. However, cache refreshing introduces additional transmissions, which can degrade the delay performance of content delivery due to the constrained bandwidth resources [1]. Therefore, efficient cache refreshing schemes should be devised to optimize both delay and content freshness. A body of works have been conducted on mobile edge caching deployment and management, but most of them focus on static content items which require no refreshing [13]. Early studies have proposed cache refreshing schemes for the database in wired networks, whereas these schemes are not applicable for mobile edge caching due to the limited bandwidth and unreliable wireless transmissions [14].

In this work, we propose a freshness-aware content refreshing scheme for cache-enabled mobile networks, in order to minimize the average service delay while guaranteeing content freshness. Age of Information (AoI) is adopted to characterize content freshness, which is the time elapsed since the generation of the current version [15]. A BS caches content items collected from source nodes to serve mobile users on demand, both through wireless transmissions. The BS always checks if the AoI of a cached item exceeds a certain threshold (defined as refreshing window) or not before delivering it to mobile users. The BS will directly deliver the cached version if still fresh. Otherwise, the BS will first fetch the latest version for refreshing, and then deliver it to mobile users.

As the system bandwidth is constrained, there exists a tradeoff between AoI and delay with respect to the refreshing window size, which is investigated in an analytical way.

Manuscript received February 12, 2020; revised September 30, 2020 and December 30, 2020; accepted March 7, 2021. Date of publication March 25, 2021; date of current version August 12, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1802803, in part by the Natural Science Foundation of China under Grant 61801011 and Grant 61902036, in part by the Beijing Municipal Natural Science Foundation under Grant L192028, and in part by the Fundamental Research Funds for the Central Universities under Grant 50100002020129001. This article was presented in part at the 2019 IEEE GLOBECOM. The associate editor coordinating the review of this article and approving it for publication was R. Tandon. *(Corresponding author: Hongbin Luo.)*

Shan Zhang, Liudi Wang, and Hongbin Luo are with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Beijing Key Laboratory of Computer Networks, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: zhangshan18@buaa.edu.cn; wangliudi@buaa.edu.cn; luohb@buaa.edu.cn).

Xiao Ma is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: maxiao18@bupt.edu.cn).

Sheng Zhou is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: sheng.zhou@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2021.3067002.

Digital Object Identifier 10.1109/TWC.2021.3067002

However, the analysis of AoI and delay is challenging due to the coupling effect of multi-content refresh and request arrival, multi-dimensional randomness of traffic arrival and wireless transmissions. To address these issues, we first study the BS service process in the single-source scenario, whereby the average AoI and service delay are derived in closed forms approximately based on queuing models and analysis. The results show that the average AoI increases with the refreshing window size in a convex manner and demonstrates an asymptotically linear relationship as the refreshing window size increases. On the contrary, the average delay is proved to decrease with the refreshing window convexly, revealing a tradeoff with the AoI. Furthermore, the proposed scheme restrains the BS from frequently refreshing an item when the request arrival rate increases, saving more bandwidth for content delivery to mobile users. Then, the results are further extended to the multi-source scenario, whereby the refreshing window is optimized for each individual content item to minimize the average delay while meeting the average AoI requirement. The problem is proved to be convex and numerical results can be obtained through MATLAB toolboxes. Extensive simulations are conducted on the OMNeT++ platform to validate the theoretical analysis. In addition, the performance of the proposed scheme is also compared with the conventional eager refreshing scheme where the items are always refreshed before delivery. The results show that the proposed scheme can effectively reduce the service delay and enhance the service capability by avoiding frequent cache refreshing, especially in heavy-loaded scenarios. In particular, the average delay can be reduced by around 80% while maintaining the AoI below one second, according to the real-trace experiments.

The main contributions of this work are as follows:

- A refreshing scheme is devised for mobile edge caching considering the dynamic variation of content information, to guarantee the freshness of user-received contents;
- The average AoI and delay performances are analyzed theoretically, revealing a tradeoff with respect to the refreshing window;
- The proposed scheme is optimized for the multi-source scenario, where the optimal refreshing window provides insights into content freshness management of practical mobile edge caching systems.

The remaining of this paper is organized as follows. The existing works on cache management are reviewed in Section II, whereby the novelty of this work is highlighted. Section III builds the system model, and the freshness-aware content refreshing scheme is proposed. The average AoI and delay are analyzed in Section IV, for the single-source scenario. Then, Section V extends the results to the multi-source scenario, based on which the refreshing window is optimized. Simulation results are provided in Section VI, followed by the conclusions in Section VII.

## II. Literature Review

Focusing on the conflict of explosive contents and constrained storage resources, extensive efforts have been denoted to enhancing the caching efficiency from the aspects of where to cache (i.e., cache deployment) [16], [17], what to cache (i.e., content placement) [18]–[22], and how to cache (i.e., cache update) [23]–[29]. Cache deployment is usually conducted in long-time scale with network planning, which determines where to deploy the cache instances under constrained budget. In this regard, the deployment costs of different network entities (like the remote servers, gateways, and heterogeneous BSs) should be considered [16], [17]. Content placement selects the items to cache based on the content popularity and user requests, aiming at different design objectives such as hit rate maximization, delay minimization, service experience enhancement, mobility support [20]–[22]. Cache update deals with the popularity variation in spatial or temporal domains. Specifically, the items with faded popularity are discarded to make room for new popular items, so as to maintain the cache hit rate [23], [24]. With perfect knowledge of content popularity, existing works have proposed to update the cache during off-peak hours, by utilizing the idle transmission resources opportunistically [25], [26]. Adaptive content update schemes remove the least used items out of cache, which can adapt to the popularity variations without perfect knowledge of content popularity [27], [28]. Furthermore, online update schemes have been also proposed based on machine learning methods such as the contextual multi-armed bandit optimization, where the BSs update the cached items based on the regular observation of content hit rate [29]. Although insightful, the-state-of-the-art studies on mobile edge caching mostly focus on static content items, wherein the content of an item is assumed to be unchanged.

Considering the dynamics of contents, the freshness or timeliness has appeared as an importance performance metric, raising the new concept of AoI [30]. The theoretical results have shown that the conventional delay-optimal transmission strategies may not be AoI-optimal [31]. Accordingly, effective transmission strategies are devised to reduce the peak or average AoI at the receiver side in different scenarios [32]–[36]. The very recent works have introduced content freshness metrics into mobile edge caching [37]–[41]. Considering that the content popularity can fade with time, Kam *et. al* have proposed to predict content popularity based on AoI and request rate [37]. However, the proposed cache update scheme still aims at maximizing the hit rate instead of maintaining content freshness [37]. Cache refreshing schemes have been devised to minimize the average AoI of a local cache system, considering the constrained transmission resources in [38] and [39]. Notice that these works focus on the AoI of cached items, whereas the freshness of user-received contents can be more important to the quality of experience. Bastopcu *et. al* have investigated the AoI of user-received contents in both single-cache and multi-cache scenarios, whereby the cache refreshing policy is devised under the predefined refreshing rate constraint [40]. Similarly, two cache refreshing schemes, namely RSUC and ReA, have been devised to minimize the service delay and the AoI of user-received content in [41]. Under the RSUC scheme, the cached items are refreshed in a round-robin manner using dedicated bandwidth, regardless of user requests. Under the ReA scheme, the cached items are refreshed with certain probability upon user requests. Although insightful, [40] and [41] do not utilize the status

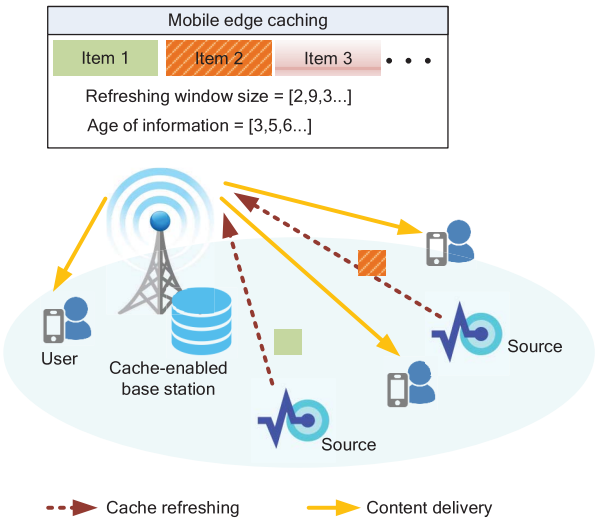


Fig. 1. Mobile edge caching with content refreshing.

information of cached items, which may cause unnecessary cache refreshing and waste transmission resources. In this work, the adopted cache refreshing scheme goes one step further, wherein the cached items are refreshed only when requested and expired.

To summarize, the novelty of this work is two-fold compared with existing works: (1) a freshness-aware content refreshing scheme is proposed and optimized to guarantee the freshness of user received contents; and (2) the interplay between the average service delay and the average AoI of user-received contents is revealed through both theoretical analysis and simulations.

## III. System Model for Content Refreshing

### A. Traffic Model

We consider a typical mobile edge caching network covered by one BS, as shown in Fig. 1. The randomly distributed source nodes monitor the surrounding environment and generate content items, such as the status of traffic jams, the arrival time of a bus, surveillance of a road crossing, availability of parking lots, and store promotions. Each source node can generate one or multiple content items. Denote by $\mathcal{C} = \{1, 2, 3, \cdots, C\}$ the set of generated content items by all source nodes, where $C = |\mathcal{C}|$. The source node keeps generating new versions of corresponding content items on demand to reflect the up-to-date information. The content generation delay is considered to be relatively small compared with transmission, which is thus ignored to simplify the analysis. The BS collects the generated content items through wireless transmission, and delivers the cached items to mobile users on demand. The total request rates of all users follow Poisson process with rate $\Lambda$. Denote by $q_c$ the probability that item-$c$ is requested, where $\sum_{c=1}^{C} q_c = 1$. Accordingly, the requests of item-$c$ also follow Poisson process of rate $\lambda_c \triangleq \Lambda q_c$.

### B. Wireless Transmission Model

Denote by $R$ the coverage radius, and $P_{\rm BS}$ the transmit power of the BS. Suppose the source nodes and users are both uniformly distributed within the coverage of BS. Considering the influence of path loss, the average service rate of content delivery is given by [42]

$$\mu_{\rm D} = \mathop{\mathbb{E}}_{r}\left[\frac{B}{L}\log_2\left(1+\frac{P_{\rm BS}r^{-\alpha}}{\sigma^2}\right)\right], \tag{1}$$

where $r$ is a random variable denoting the distance between a typical mobile user and the BS, $B$ is the available bandwidth, $L$ is the size of a content,[1] $\alpha$ is the path loss exponent of wireless transmission, $\sigma^2$ is the Gaussian noise, and the unit of $\mu_{\rm D}$ is contents per second. As the mobile users are uniformly distributed, the probability distribution of $r$ is $f_r = 2r/R^2$. Furthermore,

$$\begin{aligned}\mu_{\rm D} &= \frac{B}{L}\int_0^R \log_2\left(1+\frac{P_{\rm BS}r^{-\alpha}}{\sigma^2}\right)\frac{2r}{R^2}{\rm d}r,\\ &\gtrapprox \frac{B}{L}\int_0^R \log_2\left(\frac{P_{\rm BS}r^{-\alpha}}{\sigma^2}\right)\frac{2r}{R^2}{\rm d}r,\\ &= \frac{B}{L\ln 2}\left[\ln\frac{P_{\rm BS}}{\sigma^2}-\frac{\alpha}{2R^2}\int_0^{R^2}\ln x{\rm d}x\right]\\ &= \frac{B}{L}\log_2\frac{P_{\rm BS}\left(R/\sqrt{e}\right)^{-\alpha}}{\sigma^2}.\end{aligned} \tag{2}$$

where the symbol $\gtrapprox$ means "greater than or approximately equal to". Specifically, the approximation error goes to zero if $\frac{\sigma^2}{P_{\rm BS}r^{-\alpha}} \to 0$. Similarly, we obtain the average service rate for content refreshing:

$$\mu_{\rm R} \gtrapprox \frac{B}{L}\log_2\frac{P_{\rm Source}\left(R/\sqrt{e}\right)^{-\alpha}}{\sigma^2}, \tag{3}$$

where $P_{\rm Source}$ is the transmit power of source nodes. As the user received SNR should be high enough for reliable communications in practice systems, we can take equality in Eqs. (2) and (3) to conduct approximated analysis [42], [43]. Note that the quality of service can be guaranteed in a conservative manner.

### C. Service Model

Equipped with cache instances, the BSs can store content items and serve mobile users directly through one-hop transmission.[2] As such, the BS can save more bandwidth for content delivery with reduced remote content fetching. Meanwhile, the freshness-aware cache refreshing scheme is adopted at the BS. Specifically, a refreshing window is set for each item to suggest whether the content is fresh or not. The BS provides service via the single channel in a First-In-First-Out (FIFO) manner, and the service process is illustrated in Fig. 2. When a user raises a request for item-$c$, it will be served immediately if the channel is not occupied, and wait in a queue otherwise. In addition, the BS will always check the freshness before delivering the item. If the AoI of item-$c$ is smaller than a threshold $W_c$ (defined as the *refreshing window)*, the BS will deliver it to the user directly. Otherwise,

[1]We focus on the case of uniform content size.

[2]We focus on refreshing the cached content items for the predefined content placement, and thus assume all items can be cached at the BS. The model can be extended by allocating sufficiently small refreshing window to the uncached items.

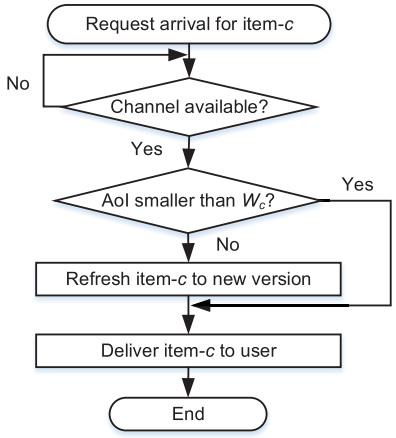


Fig. 2. Freshness-aware refreshing in mobile edge caching.

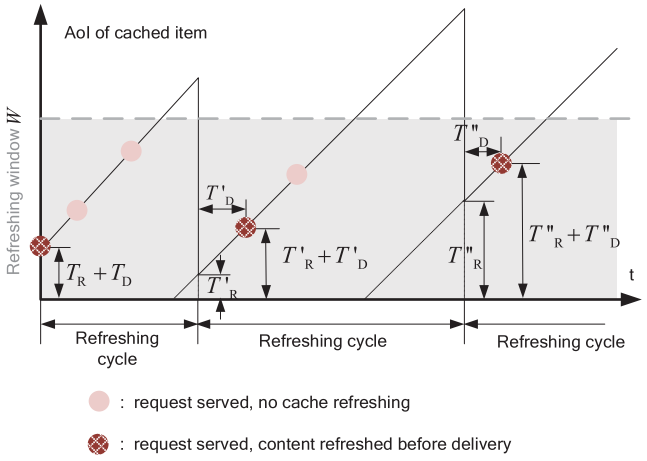


Fig. 3. AoI variation with content refreshing, single-source.

the BS will fetch the latest version of $c$ from the source node, refresh the cache and then deliver it to users.

The key design issue of the proposed scheme is to set appropriate refreshing windows. On the one hand, increasing the window size will reduce the refreshing frequency, degrading content freshness. On the other hand, decreasing the window size will introduce more transmissions due to the frequent content refreshing, degrading the service delay. Therefore, the refreshing window should be optimized to balance the content freshness and service delay. However, the analysis of AoI and delay is challenging due to the coupling effect of multi-content refreshing and request arrival, multi-dimensional randomness of traffic arrival and wireless transmissions. To address these issues, we first conduct performance analysis for the case of single source, and then extend the result to multi-source cases. The M/M/1 queue is adopted to analyze the service process in an approximated manner, which is helpful to reveal the interplay between AoI and delay with respect to the refreshing window size.

## IV. Single-Source Refreshing Analysis

To start with, we analyze the AoI and delay performances of the proposed scheme when all mobile users request the same item from one source node.

### A. Queueing Model

Denote by $T_{\mathrm{R}}$ and $T_{\mathrm{D}}$ the time consumed to refresh and deliver a content item, respectively. Assume $T_{\mathrm{R}}$ and $T_{\mathrm{D}}$ follow exponential distributions to capture the potential retransmissions considering the memoryless fading channel model [44]. Notice $\mathbb{E}[T_{\mathrm{R}}] = 1/\mu_{\mathrm{R}}$ and $\mathbb{E}[T_{\mathrm{D}}] = 1/\mu_{\mathrm{D}}$. The time consumed to serve one request is given by

$$X = \begin{cases} T_{\mathrm{D}}, & I = 0, \\ T_{\mathrm{D}} + T_{\mathrm{R}}, & I = 1, \end{cases} \tag{4}$$

where $I$ is a zero-one indicator showing whether the request is served with or without cache refreshing.

Denote by $p$ the probability that $I = 1$. The mean and variance of the service time distribution are given by

$$\mathbb{E}[X] = \frac{p}{\mu_{\mathrm{R}}} + \frac{1}{\mu_{\mathrm{D}}}, \tag{5}$$

and

$$var[X] = p^2\ var[T_{\mathrm{R}}] + var[T_{\mathrm{D}}] = \frac{1}{\mu_{\mathrm{D}}^2} + \frac{p^2}{\mu_{\mathrm{R}}^2}, \tag{6}$$

since $T_{\mathrm{D}}$ and $T_{\mathrm{R}}$ are independent. Notice that $\mathbb{E}[X]$ and $var[X]$ both depend on $p$. Thus, the refreshing probability should be derived to analyze the service delay and AoI. However, $p$ cannot be derived. The service process can be considered as a queueing system, wherein the service time is status-dependent according to Eq. (4). The challenge is that the transition of indicator $I$ does not have the Markov property due to the constant refreshing window size. Accordingly, we cannot apply the embedded Markov chain method to analyze the queue length. In addition, the cache refreshing process depends on both the request arrival and the status of the cached item, which cannot be analyzed theoretically.

To get insights, we introduce the M/M/1 queue approximation, where the customer arrival and service rates are set to $\Lambda$ and $1/\mathbb{E}[X]$, respectively. The M/M/1 queueing model can capture the additional traffic load introduced by cache refreshing, according to Eq. (5). Besides, the cache refreshing process can be analyzed for the given refreshing window size $W$, since the arrival and the departure processes of an M/M/1 queue both follow Poisson process according to the Burke's Theorem. This helps to reveal the influence of refreshing window on both the AoI and the delay performances.

### B. Content Refreshing Probability

The content refreshing process depends on the AoI of cached item, as illustrated in Fig. 3. The solid line depicts the AoI of the cached item at the BS, and each circle denotes that a user is served with requested content. In addition, the solid circles denote the requests directly served without cache refreshing, while the shadowed ones represent the requests which trigger refreshing. Suppose there is one

request served at the initial time with refreshing. Accordingly, the AoI is $T_{\mathrm{R}}+T_{\mathrm{D}}$, i.e., the overall time consumed in content fetching and delivery.[3] Then, the AoI increases with time, until another request triggers cache refreshing. The content refreshing probability can be derived according to queuing theory, given by Theorem 1.

*Theorem 1:* For the single-source case, a request triggers cache refreshing with probability

$$p=\frac{1}{(W-\frac{1}{\mu_{\mathrm{R}}})\Lambda}\left[1-e^{-(W-\frac{1}{\mu_{\mathrm{R}}})\Lambda}\right], \tag{7}$$

where $W$ is the cache refreshing window size, $\Lambda$ is the request arrival rate, $\mu_{\mathrm{R}}$ is the average service rate of content fetching. In addition, the refreshing frequency is given by

$$p\Lambda=\frac{1}{W-\frac{1}{\mu_{\mathrm{R}}}}\left[1-e^{-(W-\frac{1}{\mu_{\mathrm{R}}})\Lambda}\right]. \tag{8}$$

*Proof:* According to the Burke's Theorem, for a stable M/M/1 queueing system, the customer departure process also follows Poisson process of rate $\Lambda$. Accordingly, the inter-departure time between two requests follows exponential distribution. Define the period between two successive refreshing operations as a refreshing cycle (as shown in Fig. 3), wherein $N$ requests are served directly without cache refreshing. As 1 of $N+1$ services triggers cache refreshing within the cycle, $\mathbb{E}\left[\frac{1}{N+1}\right]$ denotes the refreshing probability on average. Furthermore, the $N$th directly served request should start service before $W$, since the cached items are refreshed if the AoI exceeds the refreshing window $W$. Denote by $T'_{\mathrm{D}}$ the delivery time of the $N$th request. Accordingly, the $N$ directly served requests should complete service within $[T_{\mathrm{R}}+T_{\mathrm{D}}, W+T'_{\mathrm{D}}]$. As $T_{\mathrm{D}}$ and $T'_{\mathrm{D}}$ are i.i.d. random variables, $N$ follows Poisson distribution of mean $(W-T_{\mathrm{R}})\Lambda$. Therefore, the refreshing probability is given by:

$$\begin{aligned} p=\mathbb{E}\left[\frac{1}{N+1}\right]&=\sum_{n=0}^{\infty}\frac{1}{n+1}\frac{\bar{N}^n}{n!}e^{-\bar{N}}\\ &=\frac{1}{\bar{N}}\sum_{n=1}^{\infty}\frac{\bar{N}^n}{n!}e^{-\bar{N}}=\frac{1-e^{-\bar{N}}}{\bar{N}}, \end{aligned} \tag{9}$$

where $\bar{N}=(W-\frac{1}{\mu_{\mathrm{R}}})\Lambda$, denoting the average number of requests directly served per refreshing cycle. Theorem 1 is thus proved. ∎

Take the first- and second-order derivatives of $p$ with respect to $\bar{N}$:

$$\frac{\partial p}{\partial \bar{N}}=-\frac{1-e^{-\bar{N}}}{\bar{N}^2}+\frac{e^{-\bar{N}}}{\bar{N}}=\frac{e^{-\bar{N}}}{\bar{N}^2}\left(\bar{N}+1-e^{\bar{N}}\right), \tag{10}$$

and

$$\frac{\partial^2 p}{\partial \bar{N}^2}=\frac{2e^{-\bar{N}}}{\bar{N}^3}\left(e^{\bar{N}}-1-\bar{N}-\frac{\bar{N}^2}{2}\right). \tag{11}$$

By taking Taylor's series of $e^x$, we can prove $e^x \geq 1+x+\frac{1}{2}x^2$, $\forall x \geq 0$. Therefore, $\frac{\partial p}{\partial \bar{N}} \leq 0$ and $\frac{\partial^2 p}{\partial \bar{N}^2} \geq 0$. Thus, $p$ is a convexly decreasing function with respect to $\bar{N}$. As $\bar{N}$ is a linear increasing function of $\Lambda$ and $W$, the refreshing probability also shows convexity with respect to $\Lambda$ and $W$. In the same way, we can prove that the content refreshing frequency $p\Lambda$ is a concave increasing function with respect to $\Lambda$ and a convexly decreasing function with respect to $W$.

The important insight of Theorem 1 is that the proposed scheme restrains the BS from frequently refreshing one item of high request rate (i.e., popular content items), since the refreshing probability $p$ decreases with $\Lambda$. However, the popular items are still refreshed more frequently under the same refreshing window size $W$. Furthermore, the refreshing probability indicates whether the item is worth to cache or not. Notice that $p \to 1$ as the $\Lambda \to 0$ or $W \to \frac{1}{\mu_{\mathrm{R}}}$. Therefore, the unpopular items with strict freshness requirements always trigger refreshing, and have no need to cache if the storage resource is constrained.

### C. Age of Information Analysis

*Theorem 2:* For the single-source case, the average AoI of user-received contents is given by

$$\bar{A}=\frac{1}{\mu_{\mathrm{D}}}+\frac{1}{2}\left(W+\frac{1}{\mu_{\mathrm{R}}}\right)-\frac{1}{2\Lambda}\left(1-e^{-\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\Lambda}\right), \tag{12}$$

under the proposed freshness-aware refreshing scheme.

*Proof:* Consider one refresh cycle wherein $N+1$ requests are served. The first request triggers cache refreshing at time zero, and the AoI of user received content equals to $T_{\mathrm{R}}+T_{\mathrm{D}}$. The other $N$ requests are served during $[T_{\mathrm{R}}+T_{\mathrm{D}}, W+T'_{\mathrm{D}}]$. As the departure process of an M/M/1 queue follows Poisson process, the $N$ requests departures uniformly in $[T_{\mathrm{R}}+T_{\mathrm{D}}, W+T'_{\mathrm{D}}]$. In addition, the AoI of user received content equals to the departure time, since the content version is generated at time zero. Therefore, the average AoI at user side is given by

$$\begin{aligned} \bar{A}&=\mathop{\mathbb{E}}_{\{N,T_{\mathrm{R}},T_{\mathrm{D}},T'_{\mathrm{D}}\}}\left[\frac{1}{N+1}\left(T_{\mathrm{R}}+T_{\mathrm{D}}+N\frac{T_{\mathrm{R}}+T_{\mathrm{D}}+W+T'_{\mathrm{D}}}{2}\right)\right]\\ &=\mathop{\mathbb{E}}_{\{N,T_{\mathrm{R}},T_{\mathrm{D}},T'_{\mathrm{D}}\}}\left[\frac{W+T_{\mathrm{R}}+T_{\mathrm{D}}+T'_{\mathrm{D}}}{2}-\frac{W-T_{\mathrm{R}}-T_{\mathrm{D}}+T'_{\mathrm{D}}}{2(N+1)}\right]\\ &=\frac{1}{\mu_{\mathrm{D}}}+\frac{1}{2}\left(W+\frac{1}{\mu_{\mathrm{R}}}\right)-\frac{1}{2}\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\mathop{\mathbb{E}}_{\{N\}}\left[\frac{1}{N+1}\right]\\ &=\frac{1}{\mu_{\mathrm{D}}}+\frac{1}{2}\left(W+\frac{1}{\mu_{\mathrm{R}}}\right)-\frac{1}{2\Lambda}\left(1-e^{-\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\Lambda}\right), \end{aligned} \tag{13}$$

Theorem 2 is thus proved. ∎

Based on Theorem 2, we further analyze the relationship between AoI and refreshing window size. Take the first- and second-order derivatives of $\bar{A}$ with respect to $W$:

$$\begin{aligned} \frac{\partial \bar{A}}{\partial W}&=\frac{1}{2}-\frac{1}{2\Lambda}e^{-\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\Lambda}\Lambda\\ &=\frac{1}{2}\left[1-e^{-\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\Lambda}\right]\geq 0, \end{aligned} \tag{14}$$

$$\frac{\partial^2 \bar{A}}{\partial W^2}=\frac{1}{2}\Lambda e^{-\left(W-\frac{1}{\mu_{\mathrm{R}}}\right)\Lambda}\geq 0. \tag{15}$$

Therefore, $\bar{A}$ is a convexly increasing function with respect to $W$. In specific, $\frac{\partial^2 \bar{A}}{\partial W^2} \to 0$ and $\frac{\partial \bar{A}}{\partial W} \to \frac{1}{2}$ as $W \to \infty$, according to Eqs. (14) and (15), respectively. Thus, $\bar{A}$ is an asymptotically linear function of $W$.

[3]We consider that the source node can always generate a new version with ignorable delay upon request.

Theorem 2 also indicates how the request arrival rate influences the content freshness. Take the first- and second-order derivatives of $\bar{A}$ with respect to $\Lambda$:

$$\begin{aligned}\frac{\partial \bar{A}}{\partial \Lambda} &= \frac{1}{2\Lambda^2} - \frac{1}{2\Lambda^2}e^{-\bar{N}} + \frac{W-\frac{1}{\mu_{\rm R}}}{2\Lambda}e^{-\bar{N}} \\ &= -\frac{e^{-\bar{N}}}{2\Lambda^2}\left[1+\bar{N}-e^{\bar{N}}\right] \geq 0, \end{aligned} \tag{16}$$

$$\begin{aligned}\frac{\partial^2 \bar{A}}{\partial \Lambda^2} &= -\left(\frac{e^{-\bar{N}}}{\Lambda^3}+\frac{(W-\frac{1}{\mu_{\rm R}})e^{-\bar{N}}}{2\Lambda^2}\right)\left(e^{\bar{N}}-\bar{N}-1\right) \\ &\quad + \frac{(W-\frac{1}{\mu_{\rm D}})e^{-\bar{N}}}{2\Lambda^2}\left(e^{\bar{N}}-1\right) \\ &= -\frac{e^{-\bar{N}}}{\Lambda^3}\left(e^{\bar{N}}-1-\bar{N}-\frac{\bar{N}^2}{2}\right) \leq 0, \end{aligned} \tag{17}$$

according to the Taylor formula, where $\bar{N} = \Lambda\left(W-\frac{1}{\mu_{\rm R}}\right)$. Therefore, the average AoI $\bar{A}$ is a concave increasing function with respect to the request rate. Two special cases are provided.

*(1) Unpopular Content:* $\Lambda \to 0$

The average AoI and content refreshing probability are given by

$$\begin{aligned}\lim_{\Lambda\to 0}\bar{A} &= \frac{1}{\mu_{\rm D}} + \frac{1}{2}\left(W+\frac{1}{\mu_{\rm R}}\right) - \frac{1}{2}\left(W-\frac{1}{\mu_{\rm R}}\right) \\ &= \frac{1}{\mu_{\rm R}} + \frac{1}{\mu_{\rm D}}, \end{aligned} \tag{18}$$

and

$$\lim_{\Lambda\to 0} p = \lim_{\bar{N}\to 0}\frac{1-e^{-\bar{N}}}{\bar{N}} = 1. \tag{19}$$

In this case, all requests are served with cache refreshing, and the mobile edge caching is merely utilized.

*(2) Popular Content:* $\Lambda \to \infty$

The average AoI and content refreshing probability are given by

$$\lim_{\Lambda\to\infty}\bar{A} = \frac{1}{\mu_{\rm D}} + \frac{1}{2}\left(W+\frac{1}{\mu_{\rm R}}\right), \tag{20}$$

and

$$\lim_{\Lambda\to\infty} p = \lim_{\bar{N}\to\infty}\frac{1-e^{-\bar{N}}}{\bar{N}} = 0. \tag{21}$$

The results indicate that almost all requests are directly served without cache refreshing. In this case, the mobile edge caching plays a key role.

### D. Service Delay Analysis

*Theorem 3:* For the single-source case, the average service delay of the proposed freshness-aware cache refreshing scheme is given by

$$\bar{D} = \frac{\frac{p}{\mu_{\rm R}}+\frac{1}{\mu_{\rm D}}}{1-\frac{p\Lambda}{\mu_{\rm R}}-\frac{\Lambda}{\mu_{\rm D}}}, \tag{22}$$

where $p$ is the refreshing probability given by Eq. (7).

*Proof:* According to the M/M/1 queueing model, the average service delay is given by $\bar{D} = \frac{1}{\frac{1}{\mathbb{E}[X]}-\Lambda}$. Substituting Eq. (5), Theorem 3 can be proved. ∎

According to (22), the average delay increases with the refreshing probability $p$. This is reasonable since the queueing delay increases with traffic load. As $p$ decreases with $W$, the average delay decreases with the refreshing window. On the other hand, the average AoI increases with $W$ according to Theorem 2. Thus, the average AoI and service delay can tradeoff by adjusting the refreshing window size. Notice that

$$\lim_{W\to\frac{1}{\mu_{\rm R}}}\bar{D} = \lim_{p\to 1}\bar{D} = \frac{1}{\frac{\mu_{\rm R}\mu_{\rm D}}{\mu_{\rm R}+\mu_{\rm D}}-\Lambda} \triangleq \bar{D}_{\max}, \tag{23}$$

$$\lim_{W\to\infty}\bar{D} = \lim_{p\to 0}\bar{D} = \frac{1}{\mu_{\rm D}-\Lambda} \triangleq \bar{D}_{\min}, \tag{24}$$

indicating how much delay can be traded by sacrificing the content freshness. Examples are illustrated to offer insights.

Case-1: $\mu_{\rm R}=\mu_{\rm D}$, $\Lambda = \frac{1}{2}\frac{1}{\frac{1}{\mu_{\rm R}}+\frac{1}{\mu_{\rm D}}}$: $\bar{D}_{\min} = \frac{1}{3}\bar{D}_{\max}$;

Case-2: $\mu_{\rm R}=\mu_{\rm D}$, $\Lambda = \frac{9}{10}\frac{1}{\frac{1}{\mu_{\rm R}}+\frac{1}{\mu_{\rm D}}}$: $\bar{D}_{\min} = \frac{1}{11}\bar{D}_{\max}$;

Case-3: $\mu_{\rm R}=\frac{1}{2}\mu_{\rm D}$, $\Lambda = \frac{1}{2}\frac{1}{\frac{1}{\mu_{\rm R}}+\frac{1}{\mu_{\rm D}}}$: $\bar{D}_{\min} = \frac{1}{5}\bar{D}_{\max}$.

Case-1 corresponds to the case that content fetching and delivery share the same transmission rates, and the system is half-loaded without mobile caching. Case-2 is more heavily loaded, but has the same channel condition of Case-1. Case-3 is half-loaded, whereas the file fetching takes longer time, corresponding to the case that source nodes use lower transmit power. These cases suggest that the proposed scheme can effectively reduce the average delay by restraining frequent cache refreshing. For example, the delay can be reduced by 60% to 90% according to the three cases. In addition, increasing refreshing window size is more beneficial on delay reduction in heavily-loaded networks with lower transmission rate of source nodes.

## V. Multi-Source Refreshing Optimization

In this section, the analytical results are further extended to the multi-source scenario, whereby the refreshing window size is optimized to minimize the average delay under the average AoI constraint of all sources.

### A. AoI and Delay Analysis

The queueing model can be extended to depict the service process of multi-source case, whereas the request arrival and service rates are different. Similarly, we study the departure process of the queue to analyze the refreshing probability of each item, whereby the average service rate can be derived. Then, the average AoI and delay can be obtained, given by Theorem 4.

*Theorem 4:* In the multi-source scenario, the average AoI of user received content item-$c$ is given by

$$\bar{A}_{{\rm M},c} = \frac{1}{\mu_{\rm D}} + \frac{1}{2}\left[W_c + \frac{1}{\mu_{\rm R}}\right] - \frac{1}{2\lambda_c}\left[1-e^{-\lambda_c\left(W_c-\frac{1}{\mu_{\rm R}}\right)}\right]. \tag{25}$$

The average service delay is given by

$$\bar{D}_{\rm M} = \frac{\frac{\bar{P}}{\mu_{\rm R}}+\frac{1}{\mu_{\rm D}}}{1-\frac{\bar{P}\Lambda}{\mu_{\rm R}}-\frac{\Lambda}{\mu_{\rm D}}}, \tag{26}$$

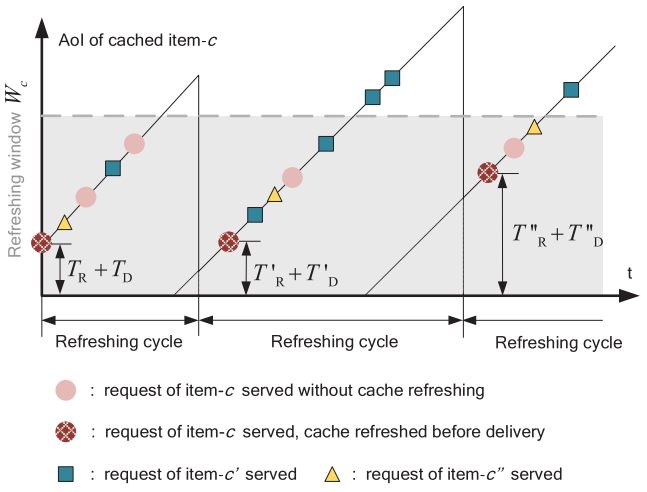


Fig. 4. AoI variation with content refreshing, multi-source.

where $\Lambda = \sum_{c=1}^{C} \lambda_c$ is the total request arrival rate of all items, $\bar{P}$ is the average cache refreshing probability

$$\bar{P} = \frac{\sum_{c=1}^{C} \lambda_c p_c}{\sum_{c=1}^{C} \lambda_c}, \tag{27}$$

and $p_c$ is the cache refreshing probability of item-$c$

$$p_c = \frac{1 - e^{-\lambda_c \left(W_c - \frac{1}{\mu_R}\right)}}{\lambda_c \left(W_c - \frac{1}{\mu_R}\right)}. \tag{28}$$

*Proof:* The BS service process is approximated as a FIFO M/M/1 queue with arrival rate of $\Lambda$. If the queue is stable, the departure process of all requests follows Poisson process, as illustrated in Fig. 4. As for item-$c$, the departure process also follows Poisson process of rate $\lambda_c$, as a randomly thinning of all requests. Therefore, the refreshing probability can be derived in the same way as the single-source scenario, given by (28). The average refreshing probability of all items can be obtained as (27), and the average service time is given by

$$\mathbb{E}[X_M] = \frac{\bar{P}}{\mu_R} + \frac{1}{\mu_D}. \tag{29}$$

Thus, the average delay is given by $\bar{D}_M = \frac{1}{\frac{1}{\mathbb{E}[X_M]} - \Lambda}$, which can be rewritten as (26). Notice that the average AoI of item-$c$ is not influenced by the service of other items if the queue is stable, and shares the same form of the single-source case as (25). Hence, Theorem 4 is proved. ∎

Theorem 4 indicates the main influencing factors on the service delay and average AoI of user-received contents. When the BS is not overloaded (i.e., stable queue), the average AoI of an item mainly depends on the refreshing window size and the request arrival rate of the corresponding item, and is merely influenced by the other items. Unlikely, the average delay depends on the intensity and refreshing window of all items, showing a coupling effect. This result indicates that the content freshness can be guaranteed by setting appropriate refreshing windows, regardless of the other items. Furthermore, the refreshing windows of each item should be jointly optimized to enhance the system-level delay performance.

### *B. Refreshing Window Optimization*

Based on the result of Theorem 4, we optimize the refreshing window to minimize the average delay while meeting the AoI requirement. The problem can be formulated as follows.

$$\min_{\{W_c\}} \ \bar{D}_M \tag{30a}$$

$$\text{(P1)} \ s.t. \ \frac{1}{\Lambda} \sum_{c=1}^{C} \lambda_c \bar{A}_{M,c} \le \hat{A}, \tag{30b}$$

$$\sum_{c=1}^{C} \lambda_c < \frac{1}{\frac{\bar{P}}{\mu_R} + \frac{1}{\mu_D}}, \tag{30c}$$

$$W_c \ge \frac{1}{\mu_R}, \quad c = 1, 2, \ldots, C, \tag{30d}$$

where $\hat{A}$ is the required average AoI, and constraint (30c) guarantees the system stability. (P1) applies to the case where contents have different popularities but require the same level of freshness (e.g., a network slice providing the same type of content service).

Denote by $\bar{N}_c = \lambda_c \left(W_c - \frac{1}{\mu_R}\right)$ for $c \in \mathcal{C}$, which reflects the average number of requests served without cache refreshing in each cycle. We rewrite (P1) in terms of $\{\bar{N}_c\}$ without (30c).

$$\min_{\{\bar{N}_c\}} \ \sum_{c=1}^{C} \frac{1 - e^{-\bar{N}_c}}{\bar{N}_c} \lambda_c \tag{31a}$$

$$\text{(P2)} \ s.t. \ \sum_{c=1}^{C} \left(\bar{N}_c + e^{-\bar{N}_c}\right) \le 2\Lambda\hat{A} + C - 2\Lambda \left(\frac{1}{\mu_R} + \frac{1}{\mu_D}\right), \tag{31b}$$

$$\bar{N}_c \ge 0, \quad c = 1, 2, \ldots, C, \tag{31c}$$

The objective function of (P2) is to minimize the average refreshing probability, which is equivalent to minimize the average delay according to Eq. (29). In addition, (30b) is equivalent to (31b), and (30d) is equivalent to (31c), according to the linear relationship between $W_c$ and $\bar{N}_c$. Although (P1) and (P2) are not mathematically equivalent, we can find the optimal solution to (P1) by solving (P2), if (P1) is feasible.

*Proposition 1:* Denote by $\{\bar{N}_c^*\}$ the optimal solution to (P2) and $W_c^* = \frac{\bar{N}_c^*}{\lambda_c} + \frac{1}{\mu_R}$ for $c \in \mathcal{C}$. $\{W_c^*\}$ is the optimal solution to (P1) if it meets (30c). Otherwise, (P1) is not feasible.

*Proof:* This proposition can be proved by contradiction. Suppose $\{W_c'\} \neq \{W_c^*\}$ the optimal solution to (P1).

If $\{W_c^*\}$ satisfies (30c), we have $\bar{D}_M|_{\{W_c'\}} < \bar{D}_M|_{\{W_c^*\}}$. Denote by $\bar{N}_c' = \lambda_c \left(W_c' - \frac{1}{\mu_R}\right)$ for $c \in \mathcal{C}$. Then, $\bar{D}_M|_{\{\bar{N}_c'\}} < \bar{D}_M|_{\{\bar{N}_c^*\}}$. Therefore, $\bar{P}|_{\{\bar{N}_c'\}} < \bar{P}|_{\{\bar{N}_c^*\}}$ according to Eqs. (26-28), which means that $\bar{N}_c'$ can lower the objective function of (P2) compared with $\bar{N}_c^*$. As $\bar{N}_c'$ is also feasible to (P2), $\bar{N}_c^*$ is not the optimal solution to (P2), which is contradictory. Hence, $\{W_c^*\}$ is the optimal solution to (P1) if $\{W_c^*\}$ satisfies (30c).

Secondly, suppose $\{W_c^*\}$ cannot satisfy (30c). As the right part of (30c) decreases with $\bar{P}$, $\bar{P}|_{\{W_c^*\}} > \bar{P}|_{\{W_c'\}}$. Accordingly, $\bar{P}|_{\{\bar{N}_c^*\}} > \bar{P}|_{\{\bar{N}_c'\}}$, which means $\{\bar{N}_c'\}$ can further

reduce the objective function of (P2). Thus, $\{\bar{N}_c^*\}$ is not optimal to (P2), which is contradictory. Therefore, (P1) has no feasible solution if $\{W_c^*\}$ cannot satisfy (30c).

Hence, Proposition 1 is proved. ∎

By taking the second-order derivative of the objective function with respect to $\bar{N}_c$, we have

$$\frac{2\lambda_c^2 e^{-\bar{N}_c}}{\bar{N}_c^3}\left(e^{\bar{N}_c}-1-\bar{N}_c-\frac{1}{2}\bar{N}_c^2\right), \tag{32}$$

which is positive for $\bar{N}_c > 0$, according to the Taylor formula. The second-order derivative of constraint (31b) with respect to $\bar{N}_c$ equals to $e^{-\bar{N}_c}$, which is also positive. As the other constraint (31c) is a linear equation of $\bar{N}_c$, problem (P2) is convex. Thus, (P2) can be addressed through convex optimization toolboxes.

We further apply the Lagrange method and analyze the optimal condition to offer insights on the setting of refreshing window. The Lagrangian function of (P2) is given by

$$\begin{aligned} &F(\bar{N}_1,\bar{N}_2,\cdots,\bar{N}_C)\\ &=\sum_{c=1}^{C}\frac{1-e^{-\bar{N}_c}}{\bar{N}_c}\lambda_c-\sum_{c=1}^{C}\nu_c\bar{N}_c\\ &\quad+\nu_0\left[\sum_{c=1}^{C}\left(\bar{N}_c+e^{-\bar{N}_c}\right)-2\Lambda\hat{A}-C+2\Lambda\left(\frac{1}{\mu_{\mathrm{R}}}+\frac{1}{\mu_{\mathrm{D}}}\right)\right], \end{aligned} \tag{33}$$

where $\nu_0$, $\nu_1$, $\cdots$, $\nu_c$ are the Lagrangian multipliers corresponding to the two constraints. By taking derivative of (33) with respect to $\bar{N}_c$, the KKT condition can be obtained:

$$\frac{\bar{N}_c e^{-\bar{N}_c}-1+e^{-\bar{N}_c}}{\bar{N}_c^2}\lambda_c+\nu_0\left(1-e^{-\bar{N}_c}\right)-\nu_c=0. \tag{34}$$

Assume $\bar{N}_c \neq 0$, and thus $\nu_c = 0$ according to the complementary slackness conditions. Thus, (34) can be written as

$$\begin{aligned} \nu_0 &= \frac{\lambda_c}{\bar{N}_c}\frac{1}{1-e^{-\bar{N}_c}}\left(e^{-\bar{N}_c}-\frac{1}{\bar{N}_c}+\frac{e^{-\bar{N}_c}}{\bar{N}_c}\right)\\ &=\frac{\lambda_c}{\bar{N}_c}\frac{1}{1-e^{-\bar{N}_c}}\left[1-\left(1-e^{-\bar{N}_c}\right)-\frac{1-e^{-\bar{N}_c}}{\bar{N}_c}\right]\\ &=\frac{\lambda_c}{\bar{N}_c}\left[\frac{1}{1-e^{-\bar{N}_c}}-1-\frac{1}{\bar{N}_c}\right]. \end{aligned} \tag{35}$$

Next, we prove that $\nu_0$ increases with $W_c$ and $\lambda_c$. Denote by $\tilde{W}_c = W_c - \frac{1}{\mu_{\mathrm{R}}} - \frac{1}{\mu_{\mathrm{D}}}$ for notation simplicity. Thus,

$$\nu_0=\frac{1}{\tilde{W}_c}\left[\frac{1}{1-e^{-\tilde{W}_c\lambda_c}}-\frac{1}{\tilde{W}_c\lambda_c}-1\right]. \tag{36}$$

Take derivative, and we have

$$\begin{aligned} &\frac{\partial\nu_0}{\partial\tilde{W}_c}\\ &=\frac{\partial\nu_0}{\partial\bar{N}_c}\frac{\partial\bar{N}_c}{\partial\tilde{W}_c}=\frac{\partial}{\partial\bar{N}_c}\left\{\frac{\lambda_c^2}{\bar{N}_c}\left[\frac{1}{1-e^{-\bar{N}_c}}-\frac{1}{\bar{N}_c}-1\right]\right\}\\ &=\lambda_c^2\left[-\frac{1}{\bar{N}_c^2\left(1-e^{-\bar{N}_c}\right)}+\frac{e^{-\bar{N}_c}}{\bar{N}_c\left(1-e^{-\bar{N}_c}\right)^2}+\frac{2}{\bar{N}_c^3}+\frac{1}{\bar{N}_c^2}\right]\\ &=\lambda_c^2\left[\frac{e^{-\bar{N}_c}}{\bar{N}_c\left(1-e^{-\bar{N}_c}\right)}\left[\frac{1}{1-e^{-\bar{N}_c}}-\frac{1}{\bar{N}_c}\right]+\frac{2}{\bar{N}_c^3}\right]. \end{aligned} \tag{37}$$

As $1-e^{-x}\leq x,\ \forall x\geq 0$, $\frac{\partial\nu_0}{\partial\tilde{W}_c}\geq 0$. Accordingly, $\nu_0$ increases with the refreshing window size $W_c$, considering the linear relationship between $W_c$ and $\tilde{W}_c$. Take derivative of $\nu_0$ with respect to $\lambda_c$,

$$\begin{aligned} \frac{\partial\nu_0}{\partial\lambda_c}&=\frac{1}{\tilde{W}_c}\left[-\frac{\tilde{W}_c e^{-\tilde{W}_c\lambda_c}}{\left(1-e^{-\tilde{W}_c\lambda_c}\right)^2}+\frac{1}{\tilde{W}_c\lambda_c^2}\right]\\ &=-\frac{e^{-\tilde{W}_c\lambda_c}}{\left(1-e^{-\tilde{W}_c\lambda_c}\right)^2}+\frac{1}{\left(\tilde{W}_c\lambda_c\right)^2}\\ &=\left(\frac{1}{\bar{N}_c}+\frac{e^{-\frac{\bar{N}_c}{2}}}{1-e^{-\bar{N}_c}}\right)\left(\frac{1}{\bar{N}_c}-\frac{1}{e^{\frac{\bar{N}}{2}}-e^{-\frac{\bar{N}_c}{2}}}\right). \end{aligned} \tag{38}$$

Denote by $Z=e^{\frac{\bar{N}}{2}}-e^{-\frac{\bar{N}_c}{2}}-\bar{N}_c$. Notice that

$$\frac{\partial Z}{\partial\bar{N}_c}=\frac{1}{2}e^{\frac{\bar{N}}{2}}+\frac{1}{2}e^{-\frac{\bar{N}}{2}}-1\geq 0. \tag{39}$$

Thus, $Z|_{\bar{N}_c>0}\geq Z|_{\bar{N}_c=0}=1-1-0=0$. Therefore, $\frac{\partial\nu_0}{\partial\lambda_c}\geq 0$, and $\nu_0$ increases with $\lambda_c$. For two items $c$ and $c'$, if $\lambda_c\geq\lambda_{c'}$, the optimal refreshing window size should satisfy $W_c < W_{c'}$ according to the optimal condition (35).

The important insight is that the popular contents should be set with a smaller refreshing window to minimize the service delay while meeting the system-level freshness requirement. Although (P1) focuses on the single-class traffic which has the same average AoI requirements, it can be easily extended with the network slicing techniques. Specifically, the transmission resources can be sliced such that each traffic class can enjoy logically isolated service. Then, constraint (30b) can be divided into multiple constraints to reflect differentiated freshness requirements. The problem is equivalent to minimize the average refreshing probability of each class, which can be decoupled as solved in the same way as (P1).

## VI. Simulation and Numerical Results

To validate the theoretical analysis, we conduct system-level simulations on the event-based OMNeT++ simulator, where the user locations, content requests and packet transmission time are generated randomly by the Monte Carlo method. The BS serves the requests in a FIFO manner and implements the proposed cache refreshing scheme. The derived refreshing probability, average AoI and delay are compared with the simulation results for both the single- and multi-source scenarios. In addition, the refreshing window is optimized in the multi-source scenario, and the influence of important system parameters is illustrated. Furthermore, the performance of the proposed scheme is also evaluated in a practical urban scenario, by implementing the Veins framework to simulate the real-trace user mobility and traffic demand [45]. Simulation parameters are listed in Table I, unless stated otherwise [42].

### A. Validation of Single-Source Analysis

To begin with, simulations are conducted in the single-source scenario, where users are uniformly distributed within coverage and randomly raise requests. In each

TABLE I
SIMULATION PARAMETERS

| Parameter | Value |
|---|---|
| coverage radius $R$ | 1000 m |
| system bandwidth $B$ | 10 MHz |
| packet size $L$ | 10 KB |
| path loss factor $\alpha$ | 4 |
| addictive white noise $\sigma^2$ | -95 dbm |
| transmit power of BS $P_{\mathrm{BS}}$ | 1 W |
| transmit power of source nodes $P_{\mathrm{Source}}$ | 0.1 W |
| request arrival rate $\Lambda$ | 1000 /s |

simulation, the BS service process is simulated under the predefined refreshing window and request arrival rate. The average AoI, delay, and refreshing probability are calculated based on 1 million request samples. In addition, the theoretical results of average AoI, delay, and refreshing probability are obtained based on Theorems 1-3, which are compared with the simulation results in Figs. 5 and 6. The simulation and theoretical results are shown to be quite close in general, validating the approximated analysis. Furthermore, the results reveal the influence of key parameters.

The influence of refreshing window is demonstrated in Fig. 5. Figure 5(a) shows the average AoI increases with the refreshing window size in an approximated linear manner as the refresh window increases. This result is consistent with the analytical results based on Eqs. (14) and (15). Figure 5(b) shows that the average delay first decreases and then levels off with the refreshing window size, which is consistent with the analysis of Eqs. (23) and (24). The reason can be explained by Fig. 5(c). As the refreshing window increases, the probability that a request triggers cache refreshing decreases, reducing the total transmission load and the average delay. When the refreshing window is sufficiently high, the cache is merely refreshed. In this case, the system transmission load achieves its minimum, and increasing refreshing window cannot further improve the delay performance.

Figures 5(a) and (b) reveal a tradeoff relationship between AoI and delay with respect to the refreshing window size. Therefore, the proposed scheme can balance the AoI and delay performance on demand of the applications, by setting appropriate refreshing window size. Furthermore, the AoI and delay performance are quite close under different request arrival rates, whereas the refreshing probability varies significantly. The important insight is that the proposed scheme can restrain the BS from frequently refreshing the same content item of high request rate. This helps to relieve traffic congestion and thus enhances the service capability of networks. More details are provided in Fig. 6.

The influence of request arrival rate is also evaluated, as shown in Fig. 6. According to Fig. 6(a), the AoI increases with the request arrival rate concavely, and finally levels off at around half of the refreshing window size. This result is consistent with Eqs. (16,17,18,21). Notice that the average AoI is mainly influenced by the refreshing window instead of the request arrival rate. Therefore, the proposed scheme can guarantee the content freshness under different load conditions, by setting appropriate refreshing window size. The reason is explained in Fig. 6(c), where the refreshing probability is

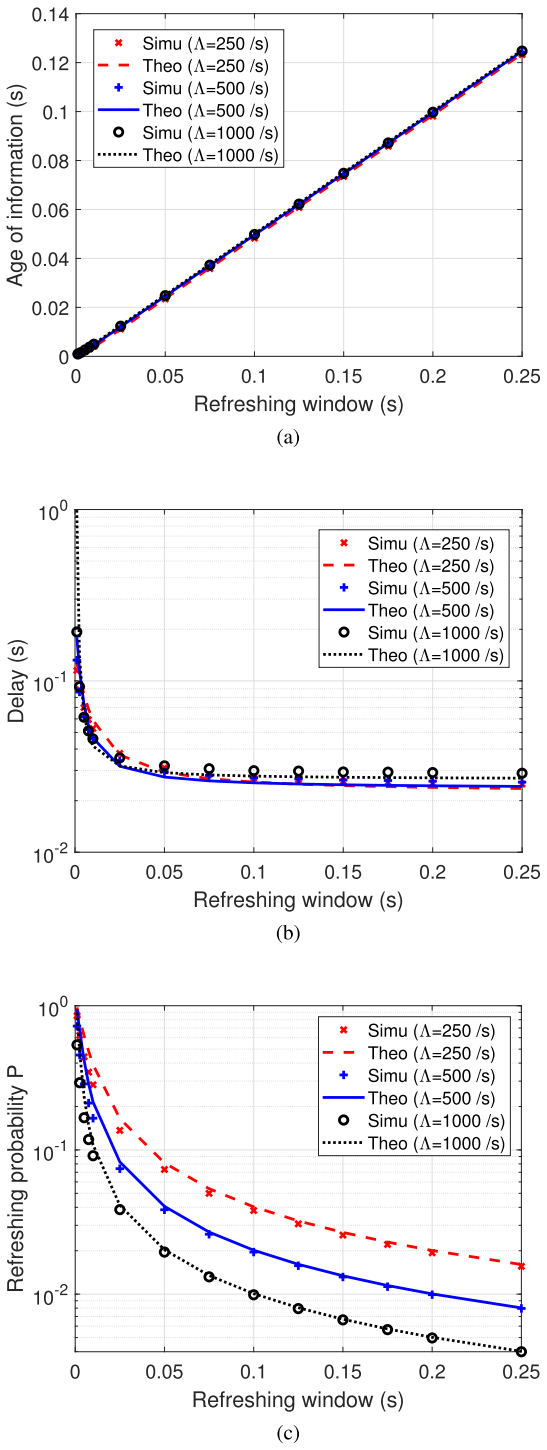


Fig. 5. Analytical results validation with respect to refreshing window in single-source scenario, (a) Age of information, (b) Delay, (c) Refresh probability.

shown to decrease with the request arrival rate. Specifically, each request triggers cache refreshing with high probability, when the request arrival rate is extremely low. In this case, the AoI mainly depends on the interval between two successive user requests rather than the refreshing window size. Unlikely, the cache is refreshed almost periodically with the cycle length equals to the refreshing window size, under the heavy traffic load. In this case, the AoI mainly depends on the refreshing window size and does not change with the request arrival rate.

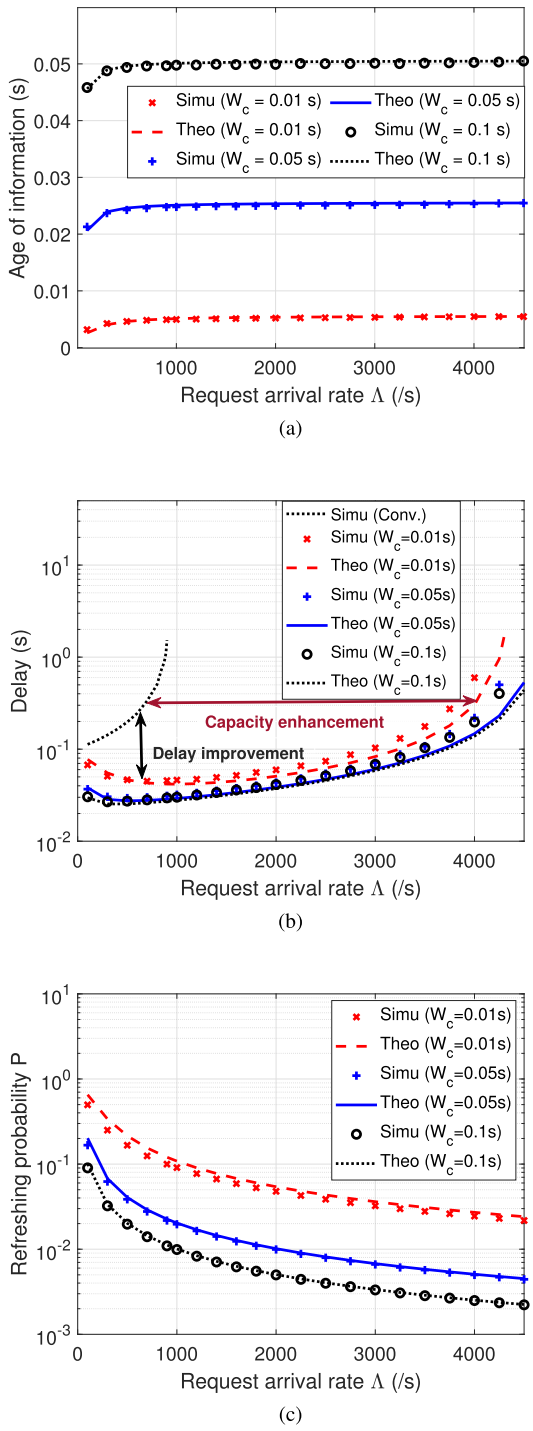


Fig. 6. Analytical results validation with respect to request arrival rate in the single-source scenario, (a) Age of information, (b) Delay, (c) Refreshing probability.

Therefore, the proposed scheme restrains the frequent cache refreshing at high request arrival rate, which is consistent with the results of Fig. 5.

The average delay is shown as a convex function of request arrival rate, which firstly decreases and then increases. This trend can be explained by Eq. (22) of Theorem 3. According to Eq. (22), the average delay increases with both the refreshing probability $p$ and request arrival rate $\Lambda$. However, the refreshing probability decreases with the request arrival rate as shown in Fig. 6(c), and thus the average delay does not always increase with the request arrival rate. The results of Fig. 6(b) indicate that the variation of delay is dominated by the refreshing probability at low request arrival rate. In addition, the average delay of conventional scheme is also illustrated as the dotted dash line, where the cache is refreshed with probability 1 (i.e., the eager refreshing scheme proposed for wired networks [14]). The results show that the proposed scheme can effectively reduce the average delay and increase the BS capacity compared with the conventional scheme. For example, if the average delay requirement is 1 second, the capacity (the maximal request arrival rate) is less than 1000 /s under the conventional scheme. In comparison, the proposed scheme can increase the capacity to 4000 /s with $W_c$ set to 0.01s, wherein the average AoI of user received content is around 5 ms. Furthermore, Fig. 6(b) shows that the capacity can be enhanced by enlarging the refreshing window but presents a marginal gain, which is consistent with Fig. 5(b).

The gap between the simulation and the analytical results mainly comes from the approximation error in analysis, i.e., using the M/M/1 queueing to approximate the service process. By approximating the service process as an M/M/1 queue, the correlation among the service time of successive requests is overlooked. As the correlation effect is more significant for higher refreshing probability, the approximation error is larger for smaller refreshing window size or lower request arrival rate. This explains why the analytical and the simulation results show a wider gap in certain regions. For example, in Fig. 6(c), the approximation error is more significant when the request arrival rate is low and $W_c = 0.01$ s. Therefore, the error of delay is also larger in this case, according to Fig. 6(b). Unlikely, the error at high request arrival rate in Fig. 6(b) is caused by two factors. Firstly, there exists approximation error in the analysis of refreshing probability. Secondly, this error is enlarged in terms of the service delay, because the queueing delay is very sensitive to the traffic load when the system is heavily loaded. The latter is the dominant factor according to the results of Figs. 6(b) and 6(c), since the error of refreshing probability is relatively small but the error of average delay is significant when the traffic load is high.

### *B. Validation of Multi-Source Analysis*

Simulations are also conducted in multi-source scenario, where the content popularity is considered to follow Zipf distribution. Without losing generality, the request probability of the $c$-th item is set to

$$q_c = \frac{1/c^{\nu}}{\sum_{s=1}^{C}(1/s^{\nu})}, \tag{40}$$

where $\nu$ is a parameter reflecting the request concentration. Content items share the same request probability $1/C$ when $\nu = 0$, i.e., uniform popularity. The typical value for video-type service is $\nu = 0.56$ [46]. Simulations are conducted in case of ten items with uniform popularity under different refreshing window and request arrival rates, and the results are shown in Fig. 7. The simulation and theoretical results are

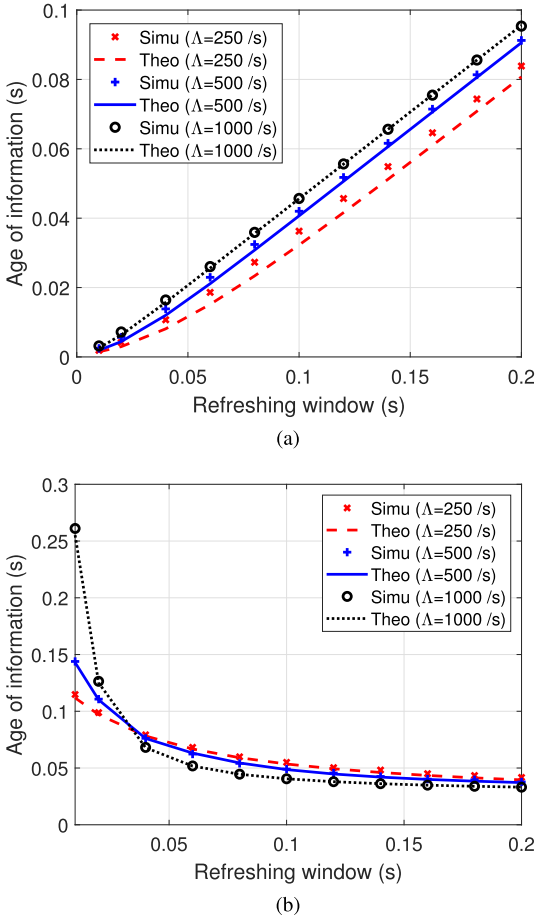


Fig. 7. Analytical results validation in the multi-source scenario with uniform popularity, (a) Age of information, (b) Delay.

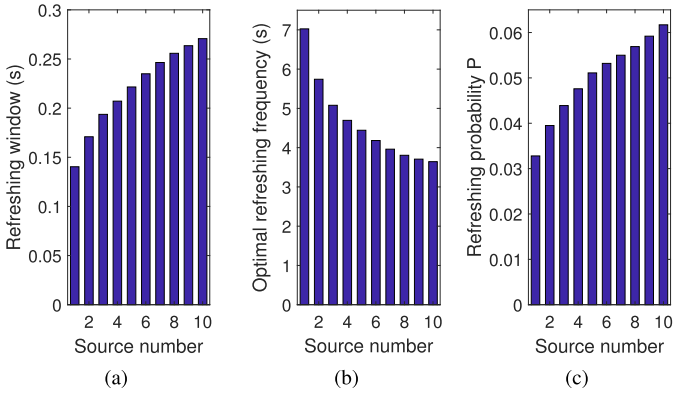


Fig. 8. Optimal refreshing of individual source: (a) Refreshing window, (b) Optimal refreshing frequency, (c) Refreshing probability, average AoI requirement 100 ms, sum request arrival rate $\Lambda = 2000$ /s, Zipf exponent $\nu = 0.56$.

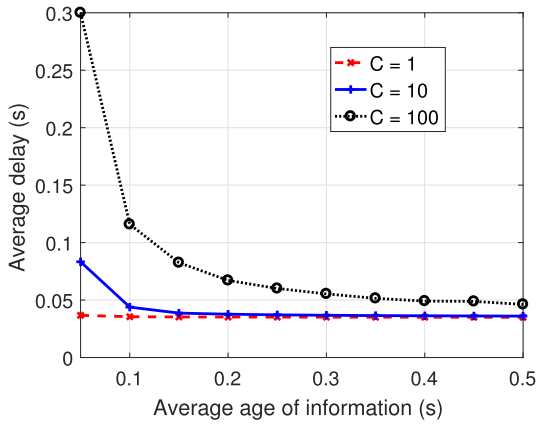


Fig. 9. Influence of item number on AoI and delay performance, sum request arrival rate $\Lambda = 2000$ /s, Zipf exponent $\nu = 0.56$.

shown to be quite close, validating the approximated analysis. Similar to the results of single-source scenario, the average AoI increases with the refreshing window in an approximately linear manner. Accordingly, the content freshness can be well guaranteed by setting appropriate refreshing window regardless of the traffic load and content popularity. In addition, the average delay decreases with the refreshing window convexly, revealing the AoI-delay tradeoff in the multi-case scenario. The reason is that cache refreshing introduces more transmissions with more source nodes, which will be discussed in details later. Similarly, the approximation error of AoI is more significant for lower request arrival rate, as shown in Fig. 7(a). In addition, the error is higher in comparison with the single-source case, due to the coupling effect among different source nodes.

### *C. Refreshing Window Optimization*

In case of nonuniform popularity, the refreshing window sizes of each content should be optimized to enhance the system-level performance. As an illustration, we consider 10 items whose popularity follows Zipf distribution of parameter 0.56, and the average AoI is required to be no larger than 100 ms. The refreshing window can be optimized by solving (P1) with MATLAB convex optimization toolbox, and the results are shown in Fig. 8. According to Fig. 8(a) and (b), the popular content items should be set with a smaller refreshing window and refreshed more frequently. However, the refreshing probability of the popular content items are rather lower according to Fig. 8(c), indicating that the proposed scheme still retains the BS from frequently refreshing popular items. In this way, the average AoI and delay is balanced.

The influences of item number and request concentration are also investigated, as illustrated in Fig. 9 and Fig. 10, respectively. The delay performance is shown to degrade as the number of source nodes $C$ increases, according to Fig. 9. The reason is that the refreshing probability of each content item is increased, introducing more transmissions. The insight is that the cost of maintaining content freshness increases with the cache size. Therefore, the cache size and content placement should be optimized considering the requirement of content freshness in practice. Furthermore, the average delay is shown to decrease with the concentration parameter $\nu$ in Fig. 10. This result indicates that mobile edge caching can provision better performance if the content requests are more concentrated, which is consistent with the case of static content items.

### *D. Real-Trace Mobility Simulations*

A simulation platform is also built based on Veins to study the influence of user mobility in practical urban scenarios,

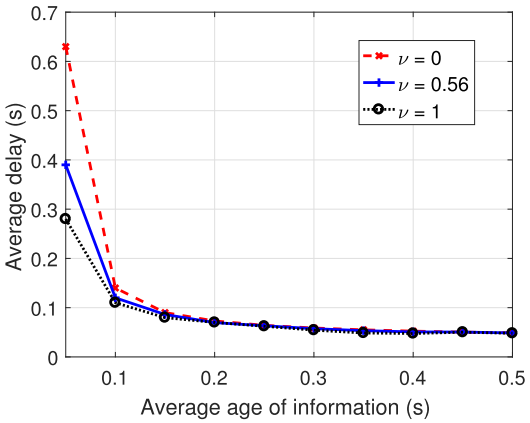


Fig. 10. Influence of content popularity distribution on AoI and delay performance, item number $C = 10$, sum request arrival rate $\Lambda = 2000$/s.

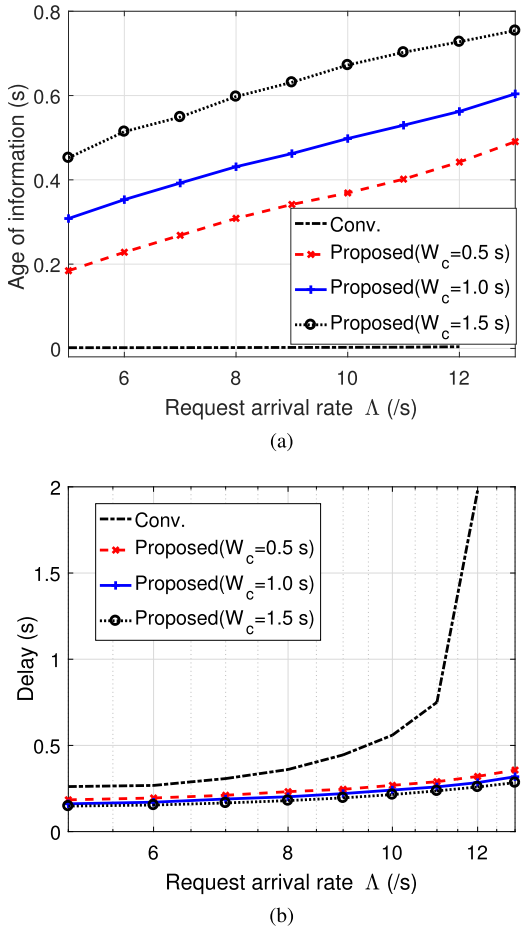


Fig. 11. Performance evaluation with mobility based on Veins, (a) AoI, (b) Delay.

using the map of Erlangen, Germany [45]. A cache-enabled BS is deployed at a road intersection, providing content service to vehicle users within 200 m. The locations of users are generated based on the Veins framework to reflect the real-trace vehicle mobility. 100 source nodes are randomly distributed within coverage, and the popularity follows Zipf distribution with $\nu = 0.56$. The BS conducts the proposed freshness-aware content refreshing scheme to balance delay and AoI performances. As comparison, the conventional scheme is adopted as a baseline, which always refreshing the cache to minimize the AoI. The simulation results of average AoI and delay are show in Fig. 11. The results show that the proposed scheme can effectively reduce the average delay compared with the conventional scheme, especially at high request arrival rates, by sacrificing the content freshness. For example, the average delay is 2 second under the conventional scheme when the request arrival rate is 12 /s. The average delay can be reduced by 85% under the proposed scheme, by setting the refreshing window to 0.5 s. Accordingly, the AoI will increase to around 420 ms, which is acceptable for many applications in practice, e.g., the traffic congestion, real-time road map, the availability of gas stations and parking lots.

## VII. Conclusion and Future Work

This work has proposed a freshness-aware content refreshing scheme for mobile edge caching systems, where the BS refreshes the cached content items based on AoI upon user requests. The average AoI and service delay have been derived in closed forms approximately, demonstrating a trade-off relationship with respect to the refreshing window size. Specifically, the average AoI of user-received content increases with the refreshing window in an asymptotically linear relationship if the system is not overloaded, while the average delay decreases in a convex manner. In the case of nonuniform content popularity, the refreshing window has been optimized for individual items to minimize the average delay while guaranteeing content freshness. Numerical results suggest to set a smaller refreshing window for popular content items, providing a guideline of the freshness-delay-optimized cache management in practice. For the future works, BSs can adopt the broadcast mode to further enhance the content delivery efficiency.

## References

[1] S. Zhang, L. Wang, H. Luo, X. Ma, and S. Zhou, "Age of information and delay tradeoff with freshness-aware mobile edge cache update," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Dec. 2019, pp. 1–6.

[2] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[3] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.

[4] X. Ma, S. Wang, S. Zhang, P. Yang, C. Lin, and X. S. Shen, "Cost-efficient resource provisioning for dynamic requests in cloud assisted mobile edge computing," *IEEE Trans. Cloud Comput.*, early access, Mar. 5, 2019, doi: 10.1109/TCC.2019.2903240.

[5] X. Cheng *et al.*, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[6] T. Yang, J. Chen, and N. Zhang, "AI-empowered maritime Internet of Things: A parallel-network-driven approach," *IEEE Netw.*, vol. 34, no. 5, pp. 54–59, Sep. 2020.

[7] L. Wang, K.-K. Wong, S. Lambotharan, A. Nallanathan, and M. Elkashlan, "Edge caching in dense heterogeneous cellular networks with massive MIMO-aided self-backhaul," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6360–6372, Sep. 2018.

[8] S. Zhang, W. Quan, J. Li, W. Shi, P. Yang, and X. Shen, "Air-ground integrated vehicular network slicing with content pushing and caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2114–2127, Sep. 2018.

[9] X. Shen *et al.*, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, Jan. 2020.

[10] Z. Gao, L. Dai, S. Han, C.-L. I, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018.
[11] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, and J. Ren, "SDN/NFV-empowered future IoV with enhanced communication, computing, and caching," *Proc. IEEE*, vol. 108, no. 2, pp. 274–291, Feb. 2020.
[12] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.
[13] P. Yang, F. Lyu, W. Wu, N. Zhang, L. Yu, and X. S. Shen, "Edge coordinated query configuration for low-latency and accurate video analytics," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4855–4864, Jul. 2020.
[14] A. Si and H. V. Leong, "Adaptive caching and refreshing in mobile databases," *Pers. Technol.*, vol. 1, no. 3, pp. 156–170, Sep. 1997.
[15] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. 8th Annu. IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw.*, Salt Lake City, UT, USA, Jun. 2011, pp. 1–9.
[16] S. Zhang, N. Zhang, P. Yang, and X. Shen, "Cost-effective cache deployment in mobile heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11264–11276, Dec. 2017.
[17] J. Kwak, Y. Kim, L. B. Le, and S. Chong, "Hybrid content caching in 5G wireless networks: Cloud versus edge caching," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3030–3045, May 2018.
[18] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.
[19] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Sep. 2016.
[20] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.
[21] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 8, pp. 1791–1805, Aug. 2018.
[22] X. Ma, A. Zhou, S. Zhang, and S. Wang, "Cooperative service caching and workload scheduling in mobile edge computing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, Jul. 2020, pp. 1–10.
[23] J. Gao, L. Zhao, and X. She, "The study of dynamic caching via state transition field—The case of time-invariant popularity," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5924–5937, Dec. 2019.
[24] J. Gao, L. Zhao, and X. Shen, "The study of dynamic caching via state transition field—The case of time-varying popularity," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5938–5951, Dec. 2019.
[25] M. Garetto, E. Leonardi, and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 1–9.
[26] H. Hsu and K.-C. Chen, "Optimal caching time for epidemic content dissemination in mobile social networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[27] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.
[28] J. Gao, S. Zhang, L. Zhao, and X. Shen, "The design of dynamic probabilistic caching with time-varying content popularity," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1672–1684, Apr. 2021.
[29] S. Muller, O. Atan, M. van der Schaar, and A. Klein, "Context-aware proactive content caching with service differentiation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1024–1036, Feb. 2017.
[30] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1–5.
[31] E. Najm and R. Nasser, "Age of information: The gamma awakening," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1–5.
[32] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form Whittle's index-enabled random access for timely status update," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1538–1551, Mar. 2020.
[33] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the Internet of Things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, Nov. 2019.
[34] Z. Jiang, S. Fu, S. Zhou, Z. Niu, S. Zhang, and S. Xu, "AI-assisted low information latency wireless networking," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 108–115, Feb. 2020.
[35] X. Chen *et al.*, "Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2268–2281, Apr. 2020.
[36] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Sep. 2016, pp. 1–8.
[37] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Information freshness and popularity in mobile caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1–5.
[38] R. D. Yates, P. Ciblat, A. Yener, and M. Wigger, "Age-optimal constrained cache updating," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1–5.
[39] J. Zhong, R. D. Yates, and E. Soljanin, "Two freshness metrics for local cache refresh," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 1–5.
[40] M. Bastopcu and S. Ulukus, "Information freshness in cache updating systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1861–1874, Mar. 2021.
[41] S. Zhang, J. Li, H. Luo, J. Gao, L. Zhao, and X. S. Shen, "Low-latency and fresh content provision in information-centric vehicular networks," *IEEE Trans. Mobile Comput.*, early access, Sep. 18, 2020, doi: 10.1109/TMC.2020.3025201.
[42] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[43] S. Zhang, N. Zhang, S. Zhou, J. Gong, Z. Niu, and X. Shen, "Energy-aware traffic offloading for green heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1116–1129, May 2016.
[44] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4196–4209, Aug. 2013.
[45] C. Sommer, R. German, and F. Dressler, "Bidirectionally coupled network and road traffic simulation for improved IVC analysis," *IEEE Trans. Mobile Comput.*, vol. 10, no. 1, pp. 3–15, Jan. 2011.
[46] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas. (IMC)*, San Diego, CA, USA, Oct. 2007, pp. 1–14.

**Shan Zhang** (Member, IEEE) received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. She was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, from 2016 to 2017. She is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing. Her research interests include mobile edge computing, wireless network virtualization, and intelligent management. She received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013. She serves as an Associate Editor for IEEE INTERNET OF THINGS JOURNAL, and *Peer-to-Peer Networking and Applications*.

**Liudi Wang** (Student Member, IEEE) received the B.S. degree from the Ocean University of China, Qingdao, China, in 2019. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Beihang University. Her research interests include mobile edge caching and wireless network communications.

**Hongbin Luo** (Member, IEEE) received the B.S. degree from Beihang University in 1999 and the M.S. (Hons.) and Ph.D. degrees in communications and information science from the University of Electronic Science and Technology of China (UESTC), in June 2004 and March 2007, respectively.

From June 2007 to March 2017, he worked with the School of Electronic and Information Engineering, Beijing Jiaotong University. From September 2009 to September 2010, he was a Visiting Scholar with the Department of Computer Science, Purdue University. He is currently a Professor with the School of Computer Science and Engineering, Beihang University. He has authored more than 50 peer-reviewed papers in leading journals, such as IEEE/ACM TRANSACTIONS ON NETWORKING and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and conference proceedings. His research interests are in the wide areas of network technologies, including network architecture and routing and traffic engineering. In 2014, he received the National Science Fund for Excellent Young Scholars from the National Natural Science Foundation of China (NSFC).

**Xiao Ma** (Member, IEEE) received the B.S. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2013, and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, in 2018.

From October 2016 to April 2017, she visited the Department of Electrical and Computer Engineering, University of Waterloo, Canada. She is currently a Lecturer with the State Key Laboratory of Networking and Switching Technology, BUPT. Her research interests include mobile cloud computing and mobile edge computing.

**Sheng Zhou** (Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2005 and 2011, respectively. In 2010, he was a Visiting Student with the Wireless System Lab, Department of Electrical Engineering, Stanford University, Stanford, CA, USA. From 2014 to 2015, he was a Visiting Researcher with the Central Research Laboratory, Hitachi Ltd., Japan. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include cross-layer design for multiple antenna systems, mobile edge computing, vehicular networks, and green wireless communications. He received the IEEE ComSoc Asia–Pacific Board Outstanding Young Researcher Award in 2017 and the IEEE ComSoc Wireless Communications Technical Committee (WTC) Outstanding Young Researcher Award in 2020.