This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

OARE: Overhearing-Aided Redundancy Elimination in Multi-Rate WLANs

Gwangwoo Park, Youngjun Shim, Insun Jang, Sangheon Pack, *Senior Member, IEEE*, and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—In general, an access point (AP) in wireless local area networks (WLANs) can be easily congested since all related data traffic goes through it. To satisfy the quality of service (QoS) requirement for its customers, protocol-independent redundancy elimination (RE) technique can effectively mitigate the congestion at the AP by reducing redundant data traffic. In this paper, we propose an overhearing-aided redundancy elimination (OARE) scheme for multi-rate WLANs by leveraging the broadcast nature of wireless medium. In OARE, the AP estimates overhearing probabilities of partially sliced chunks from an incoming packet for a destination mobile terminal by considering adaptive modulation and coding (AMC) technique. After that, the original packet is encoded by means of RE to minimize its transfer time to the destination depending on the estimated overhearing probabilities. Extensive simulation results demonstrate that OARE can reduce the packet transfer time by 7.8–23.2% compared with non-RE (or non-encoded) transmissions.**

***Index Terms*—Redundancy elimination, overhearing, broadcast, multi-rate, transfer time.**

I. Introduction

Mobile data traffic has been explosively increased due to the proliferation of mobile devices such as smartphones and tablet PCs [2]. Such a rapid increase of mobile data traffic may cause severe network congestion in wireless networks. In particular, since all data traffic is concentrated at an access point (AP) in wireless local area networks (WLANs), the AP can be heavily congested and it is hard to satisfy quality of service (QoS) for its customers. To address this problem, diverse mobile data offloading technologies have been actively discussed in the literature [3]–[6]. A typical mobile data offloading redirects mobile data traffic originally targeted for cellular networks through complementary access networks such as wireless fidelity (WiFi) hotspots or femtocells. Furthermore, mobile data offloading can be achieved by reducing duplicate data transmissions since popular contents are repeatedly requested in the networks [7]–[11]. Especially, protocol-independent traffic redundancy elimination (RE) technique has been introduced for reducing redundant data traffic at the packet-level [12]–[23].

In RE, a sender first evaluates real byte-strings of partially sliced chunks from an incoming packet in order to identify redundant byte-strings already stored at the receiver [12,13]. If the receiver has identical (or redundant) chunks to those of the sender, the sender replaces them with small meta data (i.e., encoding process) and transmits the encoded packet rather than the original one. After receiving the encoded packet, the receiver reconstructs them by replacing the meta data with the original chunks (i.e., decoding process) and passes the decoded packet to the upper layer. To this end, cache tables of the sender and the receiver should be always synchronized in order to avoid any additional cost and delay during both encoding and decoding processes. It was reported that RE can achieve the average bandwidth saving of 15–60% for access networks [14]. Recently, extensive researches for deploying RE on end hosts [15], wide-area networks [16,17], and cloud computing [18,19] have been conducted.

The performance of RE in wireless networks can be further improved by leveraging the broadcast nature of wireless medium. The transmission between a transmitter-receiver pair in wireless networks is physically broadcast and thus, neighboring mobile terminals (MTs) can obtain and store these packets without any effort. In this context, many studies have been conducted to improve the performance of wireless networks by utilizing the overheard packets [9,24]. In RE, the AP can encode the packet by referring overheard packets by the MT. However, since the AP cannot assure which packets are successfully overheard by the MT, additional techniques should be devised to determine whether the MT obtained the packet or not in the past. In Wireless Memory [20], the overhearing probability is determined by comparing the transmission rate of the destination MT and the available transmission rates of the rest MTs. It is assumed that the transmission rate of the MT reflects the channel condition of it. Therefore, if a tagged MT can use a higher (or equal) transmission rate than that of the destination MT, the AP considers the transmitted packet to be overheard by the tagged MT and adds the chunks of the packet to the cache table for the tagged MT. However, since the AP manages separate cache tables for each MT, redundant data can be stored for different MTs. On the other hand, overheard chunks are statistically determined in REfactor [21]. To this end, the fraction of nodes who overheard the packets is measured beforehand in

This work was supported in part by National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014R1A2A1A12066986) and in part by ICT R&D program of MSIP/IITP (2014-044-006-004, Next Generation WLAN System with High Efficient Performance).

A preliminary version of this paper was presented at AsiaFI 2013 Summer School, Hong Kong, China, August 2013 [1].

G. Park, Y. Shim, I. Jang, and S. Pack are with the School of Electrical Engineering, Korea University, Seoul, Korea. (e-mail:{kwoo17, yjshim, zerantoss, shpack}@korea.ac.kr)

X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. (e-mail:xshen@bbcr.uwaterloo.ca)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

a wireless mesh network testbed. However, this approach is inappropriate when the channel condition varies rapidly.

In this paper, we propose an overhearing-aided redundancy elimination (OARE) scheme to mitigate redundant data traffic load by means of RE in multi-rate WLANs. In particular, OARE exploits the broadcast nature of wireless medium to further improve the performance of RE. Moreover, since the AP has one integrated cache table instead of separate cache tables for all of MTs, redundant data between the separate tables can be significantly reduced in OARE. To calculate the overhearing probability of each chunk, the AP stores additional information such as the transmission mode and the time stamp of the chunk into the cache table. Based on this information, the chunk overhearing probability for a certain MT can be calculated through the adaptive modulation and coding (AMC) and finite-state Markov chain (FSMC) models. Therefore, the AP does not need to track any information of MTs. After obtaining the overhearing probability of each chunk in an incoming packet, the AP computes the optimal encoding threshold on the overhearing probability, denoted by δ, to minimize the expected transfer time of the packet. Since each packet has a different number of redundant chunks and their overhearing probabilities are also different, δ is configured by a per-packet basis. Finally, only chunks whose overhearing probabilities are greater than or equal to δ are encoded by replacing the corresponding fingerprints and then, the encoded packet is transmitted to the destination MT. Extensive simulation results demonstrate that OARE can reduce the packet transfer time by 7.8–23.2% compared with non-RE (or non-encoded) transmissions.

Main contribution of this paper is three-fold: 1) OARE is based on the well-defined analytical model, which can estimate the overhearing probability of the previously transmitted packet only by employing limited information such as transmission mode and elapsed time. Moreover, the analytical model is flexible so that it can be utilized for other types of networks such as opportunistic networks and wireless mesh networks; 2) we devise a comprehensive method to determine the optimal encoding threshold δ for each packet with various sizes and different numbers of redundant chunks; and 3) we provide extensive simulation results to show the effect of OARE compared with other RE schemes.

The rest of this paper is organized as follows. In Section II, the background and related works of RE are summarized. Section III describes the system model of OARE. Section IV presents the analytical model to estimate the overhearing probability and compute δ. In addition, the basic operation of OARE is described. Section V illustrates simulation results in different environments. Finally, Section VI concludes this paper.

II. Background & Related Works

Some of Internet contents are highly popular and transferred repeatedly across the network. Therefore it is possible to mitigate the network congestion and reduce the download latency by reducing duplicate data transmissions. Web cache is the most widely used method to eliminate traffic redundancy [7]–[11]. It can store and serve popular contents on behalf of contents providers. However, it has an inherent problem with identifying contents by their names, i.e., uniform resource locator (URL). Therefore, the same or partially common contents with different URLs should be delivered repeatedly, even if the same or partially common contents already exist. To remedy the limitation of Web cache, RE has been recently introduced. Unlike caching, repeated chunks can be detected by comparing real byte-strings obtained from the packet. That is, a sender can transmit only unique parts in the original packet after removing duplicate parts that a receiver already has (i.e., repeated chunks). Therefore, duplicate traffic can be significantly reduced in spite of different URLs.

Basic operation of RE consists of four blocks: 1) fingerprinting, 2) chunk matching, 3) encoding, and 4) decoding. Each block can be described as follows.

1) Fingerprinting: The representative set of fingerprints are generated from incoming packets because it is impractical to store and compare every fingerprint of each packet. To improve the performance of RE, the selected representative fingerprints should be content-based and uniformly distributed over the packets [13]. To this end, several methods such as MODP [13], MAXP [14], and SAMPLEBYTE [15], are proposed in the literature.
2) Chunk matching: The duplicate chunks of the representative fingerprints are retrieved from the cache table. There are two types of chunk matching mechanism: Chunk-Matching and Max-Matching. The same size chunks are retrieved in the packet cache in Chunk-Matching, whereas the matched region can be expanded back and forth of the original matched chunk in Max-Matching. Max-Matching usually achieves more traffic reduction than Chunk-Matching. However, it requires numerous byte comparisons to expand the matched region.
3) Encoding: The selected chunks by chunk matching are replaced with small meta data. The content of meta data can be different in diverse RE systems. For example, it consists of the address of the chunk in the cache table, length, and offset in the original packet. In addition, the corresponding fingerprint can be the meta data itself.
4) Decoding: After receiving the encoded packet, the receiver decodes the packet by replacing the meta data with the original chunks. If the decoding is failed, the receiver can request the original packet to the sender.

In the literature, diverse RE systems have been introduced to improve the performance of RE along with different network characteristics [12]. In [13] and [14], middleboxes for the RE are located at the access link between two different networks. It is revealed that 2–76% traffic is redundant depending on different protocols and 15–60% traffic in access networks can be reduced based on real Internet traffic traces. On the other hand, Aggarwal *et al.* [15] employ RE between end devices to maximize the single link traffic reduction. Specifically, to overcome the limited processing power of end devices, a new fingerprinting method with a simple cache architecture, SAMPLEBYTE, is proposed, which has lower complexity than MODP [13] and MAXP [14].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

In addition, researches for supporting RE as a network-wide service have been studied. In [16], Anand *et al.* develop redundancy-aware intra- and inter-domain routing algorithms and reduce intra- and inter-domain link loads by 10–50%. In addition, they present a network-wide RE architecture, SmartRE [17], which consists of three key elements: ingress nodes, interior nodes, and a central configuration module. Ingress nodes encode packets, whereas interior nodes only decode the packets. The encoded region of each packet can be expanded by leveraging the cached data on its route. The configuration module computes encoding and caching manifests and delivers them to ingress and interior nodes, respectively.

PACK [18] and CoRE [19] are proposed to reduce the bandwidth cost of data transfer from cloud computing. In PACK, cloud-server RE effort is offloaded to a large group of clients. It employs a receiver-based approach in which clients use newly received chunks to identify previously received chunk chains and send the prediction to the sender (i.e., server). On the other hand, in CoRE, a sender and receiver cooperative solution is proposed to simultaneously remove short-term and long-term redundancy. It has two-layer RE modules: 1) the first-layer module detects long-term redundancy by a prediction-based approach like PACK. 2) the second-layer module detects short-term redundancy by maintaining a temporary small local cache.

Recently, RE systems for wireless/mobile networks also have been actively studied. In contrast with wired networks, high packet error rate and scarcity of bandwidth are still challenging issues in wireless/mobile network. Lumezanu *et al.* suggest an informing procedure of packet loss to eliminate misleading information [22]. In Celleration [23], the cellular gateway observes the forwarded chunks to identify previously observed chunk chains similar to PACK [18]. By this approach, it is found that 46% cellular web traffic is redundant.

Moreover, the performance of RE in wireless networks can be further improved by leveraging the broadcast nature of wireless medium. Since a large group of MTs can overhear packet transmissions destined to others, the AP can exploit the chunks of the overheard packets to encode the next incoming packets. However, since the AP does not receive any acknowledgement of the overheard packets, the AP cannot assure that chunks of an incoming packet were really overheard. To remedy this problem, in Wireless Memory [20], the AP assumes that MTs can always overhear the packets which are sent with lower transmission rates than the MTs' available transmission rates. The AP should maintain separate cache tables for different MTs and the chunks of the overheard packets are stored in different cache tables repeatedly. On the other hand, in REfactor [21], the AP maintains one cache table and each stored chunk has a reception probability vector for each MT. However, the reception probability is computed heuristically based on the training data obtained from a wireless mesh network testbed. Consequently, those values are restricted to a certain network condition, and thus it is needed to compute the reception probability under more general models.

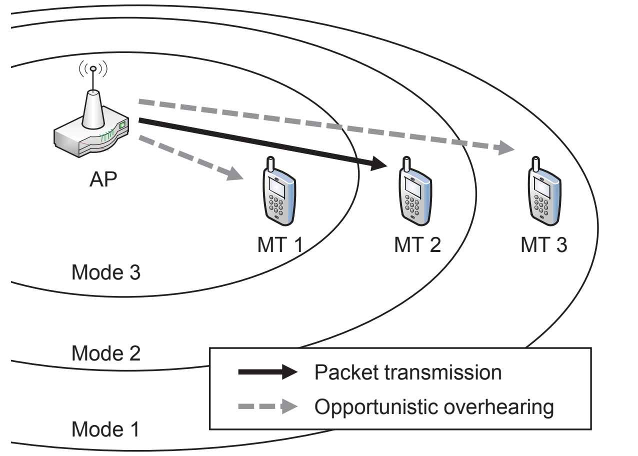


Fig. 1: Network model of OARE.

III. System Model

In this section, the system model of OARE is described. As shown in Figure 1, we consider downlink transmissions in multi-rate WLANs, in which the AP transmits a packet to the destination MT (i.e., MT 2). It is assumed that MTs stay inside a WLAN for a long time [6] and thus, the mobility of the MTs is not considered in this work. Due to the broadcast nature of wireless medium, neighbors (i.e., MT 1 and MT 3) can opportunistically obtain (or overhear) the transmitted packet depending on their own channel conditions. If the MTs overhear the packet successfully, they store a set of chunks of the overheard packet in their caches, and therefore the AP can exploit these chunks to encode the next packets destined to them (i.e., MT 1 and MT 3). However, since the packet is opportunistically obtained, the AP should first identify which redundant chunks are successfully obtained by the intended MT. To this end, the AP in OARE maintains the minimal information about the transmitted chunks in its cache and determines whether the chunk is overheard or not by computing the overhearing probability of the chunk in an analytical manner, which will be elaborated in Section IV. Consequently, OARE can provide improved scalability compared with other techniques such as Wireless Memory [20] and REfactor [21].

The following Section III-A illustrates the configuration of cache tables of AP and MT to support RE in multi-rate wireless networks. After that, Section III-B describes the wireless channel model.

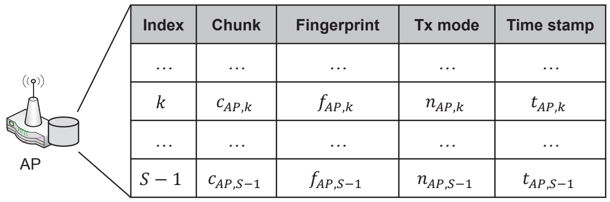
A. Cache architecture

Figure 2(a) shows the cache table of the AP that can store S items. Each item consists of four tuples as

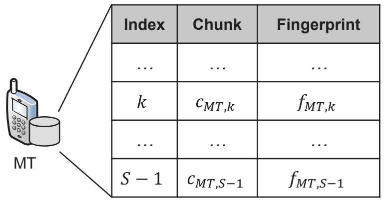
$$V_{AP,k} = \{c_{AP,k}, f_{AP,k}, n_{AP,k}, t_{AP,k}\} \tag{1}$$

where k is an item index $(0 \leq k \leq S-1)$. $c_{AP,k}$ denotes a chunk which is a byte-string for identifying the redundant region of a packet. It is reported that 64 bytes chunk is the most effective to identify the redundant data [13] [14]. $f_{AP,k}$ is a fingerprint of the corresponding chunk $c_{AP,k}$, which is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology



(a) Cache table at AP.



(b) Cache table at MT.

Fig. 2: Cache architectures in BFRE.

calculated from a hash function and is smaller than $c_{AP,k}$ (e.g., 160 bits for SHA-1). $n_{AP,k}$ denotes the transmission mode identification (ID) of $c_{AP,k}$. It is assumed that there are N transmission modes depending on wireless channel conditions and one appropriate mode $n_{AP,k}$ $(1 \leq n_{AP,k} \leq N)$ for transmitting $c_{AP,k}$ is selected by the AP. The details of the adaptive transmission mode will be described in subsection III-B. $t_{AP,k}$ indicates the time stamp when the chunk $c_{AP,k}$ was transmitted. If the current time is τ, the time stamp of every chunk in the cache table should be smaller than τ (i.e., $t_{AP,k} < \tau$) since each chunk was transmitted in the past. Both of $n_{AP,k}$ and $t_{AP,k}$ are used to compute the overhearing probability of $c_{AP,k}$ and the details of computation will be demonstrated in subsection IV-A.

To reduce the search time, a part of fingerprint is used as the index of the corresponding chunk in the cache table. The size of a chunk and a fingerprint in RE is very small (e.g., 64 and 20 bytes, respectively), and thus a number of items can be stored in the cache table. Since it requires long delay to search an arbitrary value from the large table, the first $log_2 S$ bits of the fingerprint are exploited as the index of the corresponding item in OARE and the search time is then fixed as $O(1)$. For example, when the AP searches for the redundant chunk c_{Red} in its cache table, it first calculates the corresponding fingerprint f_{Red} and computes the index ω from the first $log_2 S$ bits of f_{Red}. Then, the AP compares f_{Red} and $f_{AP,\omega}$ that is stored in the cache table. In a similar way, when a new item $V_{New} = \{c_{New}, f_{New}, n_{New}, t_{New}\}$ is added to the cache table, the AP first checks whether the index ω (which is computed from f_{New}) of the cache table is occupied or not. If it is occupied, the old one (denoted by $V_{AP,\omega}$) is replaced with the new one (i.e., V_{New}) because the new one has higher priority than the old one as least recently used (LRU) algorithm.

Figure 2(b) shows the cache table of the MT. The cache table has S items and each item $V_{MT,k}$ consists of two tuples as

$$V_{MT,k} = \{c_{MT,k}, f_{MT,k}\} \tag{2}$$

TABLE I: Transmission modes from [25].

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Modulation	BPSK	QPSK	QPSK	16-QAM	16-QAM	64-QAM
Coding rate	1/2	1/2	3/4	9/16	3/4	3/4
bits/symbol	0.50	1.00	1.50	2.25	3.00	4.50
a_n	274.72	90.25	67.62	50.12	53.40	35.35
g_n	7.99	3.50	1.69	0.66	0.38	0.09
γ_{pn}(dB)	-1.53	1.09	3.98	7.70	10.25	15.98

where $c_{MT,k}$ and $f_{MT,k}$ denote the chunk and the fingerprint, respectively. Since the role of the MTs is limited to the decoding process, the cache table of the MTs does not need a transmission mode and a time stamp of each chunk. In the same manner as that of the AP, the first $log_2 S$ bits of fingerprint is used to indicate the position of the corresponding item in the cache table. For example, when the MT receives an encoded packet and the packet has the fingerprint f_{Red}, the MT computes an index ω from the first $log_2 S$ bits of f_{Red}. Then, it replaces f_{Red} in the encoded packet with the corresponding chunk $c_{MT,\omega}$ in the cache table. The details of the decoding process will be given in subsection IV-E.

B. *Wireless channel model*

AMC or link adaptation technique is generally supported to choose the most appropriate transmission rate among multiple transmission modes when the channel condition of an intended MT is given. In Figure 1, the AP obtains the channel condition of the destination MT (i.e., MT 2) and transmits a packet with the appropriate transmission mode (i.e., Mode 2). Since a higher transmission rate (i.e., higher transmission mode) requires better channel condition to successfully transmit a packet, MT 1 has better channel condition than MT 2 while MT 3 has worse channel condition than MT 2.

In Figure 1, when the AP transmits a packet to MT 2, other MTs (i.e., MT 1 and MT 3) can also obtain the packet depending on their packet reception probabilities. To calculate the reception probability of the packet, we consider a slowly varying wireless channel [25]. The channel is frequency-flat, and remains invariant per packet while it is allowed to vary from packet to packet. For flat fading channels, the channel quality can be captured by the received signal-to-noise ratio (SNR). The received SNR can be measured at the receiver and exposed the channel state synthetically (including distance, obstacle, etc.). Since the channel varies from packet to packet, the Nakagami-m model is adopted to describe statistically [26], which is a general model to assemble different fading channels and can be represented by FSMC. The received SNR γ per packet is a random variable with a Gamma probability density function (PDF)

$$f_{\bar{\gamma}}(\gamma) = \frac{m^m \gamma^{m-1}}{\bar{\gamma}^m \Gamma(m)} \exp\left(-\frac{m\gamma}{\bar{\gamma}}\right) \tag{3}$$

where $\bar{\gamma}$ is the average SNR, m is the Nakagami fading parameter ($m \geq 0.5$), and $\Gamma(m)$ is the Gamma function.

In multi-rate wireless networks, each mode consists of a specific modulation and forward error coding (FEC) code pair as in IEEE 802.11a [27] which is listed in Table I.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

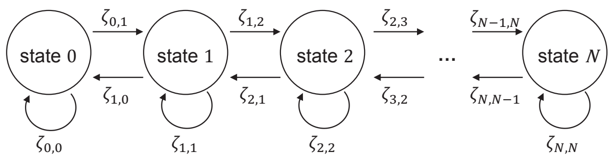


Fig. 3: FSMC state transition diagram.

Then, the received the SNR at the receiver can be divided into multiple non-overlapping intervals by thresholds γ_n $(n \in \{0, 1, ..., N+1\})$, where $\gamma_0 = 0 < \gamma_1 < \gamma_2 < ... < \gamma_{N+1} = \infty$. These thresholds correspond to the required SNR at the receiver. That is, if the received SNR γ is $[\gamma_n, \gamma_{n+1})$, transmission mode n is selected. Note that no packet is transmitted when $n = 0$ (i.e., $\gamma < \gamma_1$) since the channel is deep fading and the packet is frequently corrupted.

When the MT has the average SNR $\bar{\gamma}$, the probability that the MT uses the transmission mode n is computed as

$$\pi\left(n, \bar{\gamma}\right) = \int_{\gamma_n}^{\gamma_{n+1}} f_{\bar{\gamma}}\left(\gamma\right) d\gamma = \frac{\Gamma\left(m, \frac{m\gamma_n}{\bar{\gamma}}\right) - \Gamma\left(m, \frac{m\gamma_{n+1}}{\bar{\gamma}}\right)}{\Gamma\left(m\right)}$$

where $\Gamma\left(m, x\right)$ is a complementary incomplete Gamma function.

Let $\mathrm{PER}(n, \gamma)$ be the packet error rate (PER) when the packet is transmitted with the transmission mode n and the received SNR of the MT is γ. From [25], $\mathrm{PER}(n, \gamma)$ can be approximated as

$$\mathrm{PER}\left(n, \gamma\right) \approx \begin{cases} 1, & \text{if } 0 < \gamma < \gamma_{pn} \\ a_n \exp\left(-g_n \gamma\right), & \text{if } \gamma_{pn} \le \gamma \end{cases} \tag{4}$$

where a_n, g_n, and γ_{pn} are mode-dependent parameters and obtained by fitting (4) to the exact PER [25]. These mode-dependent parameters are summarized in Table I[1].

On one hand, in the overhearing case, transmission modes of the overheard MT and the destined MT can be different (e.g., The transmission modes of MT 1 and MT 2 are different in Figure 1.), since their channel conditions are dissimilar. Assume that the packet is transmitted with the transmission mode n (i.e., The original destination MT uses the transmission mode n.) and the overheard MT uses another transmission mode ν. If the average received SNR of the overheard MT is $\bar{\gamma}$, the average PER of the overheard packet can be computed as (5) where $b_n := m/\bar{\gamma} + g_n$.

One state in the FSMC corresponds to one transmission mode in multi-rate wireless networks [28,29] as shown in Figure 3. Let $\zeta_{i,j}$ be the state transition probability from state i to state j. Note that no packet is transmitted in state 0 to avoid frequent packet losses due to the deep fading channel. In this model, the received SNR remains at a certain level for the time duration to transmit a packet and this time duration is referred as one time slot. In addition, the state transition occurs only between adjacent states during a time slot. That is, the probability of the state transition exceeding two consecutive states is zero, i.e.,

$$\zeta_{i,j} = 0, \quad \forall \ |i - j| > 1.$$

[1]Note that the values of γ_{pn} in Table I are expressed in decibel.

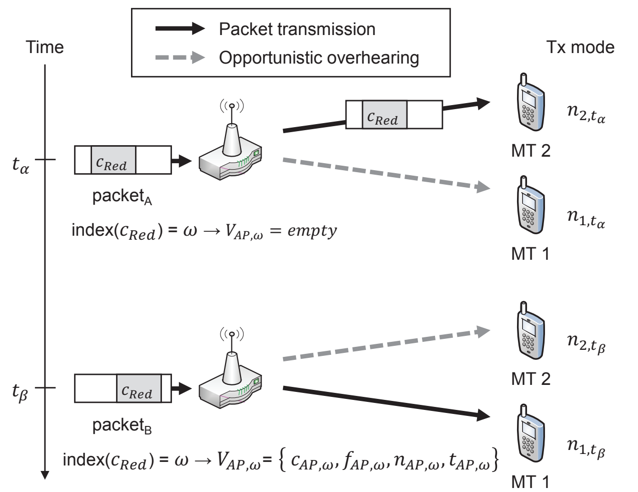


Fig. 4: An example scenario for estimating the overhearing probability. At time slot t_β, the AP has $packet_b$ to transmit to MT 1. The packet contains redundant chunk c_{Red} which was transmitted at time slot t_α and stored in the cache.

Now, the adjacent-state transition probability can be described by

$$\begin{array}{ll} \zeta_{n,n+1} = \frac{R_{n+1} T_s}{\pi(n, \bar{\gamma})}, & \text{if } n = 0, ..., N-1 \\ \zeta_{n,n-1} = \frac{R_n T_s}{\pi(n, \bar{\gamma})}, & \text{if } n = 1, ..., N \end{array} \tag{6}$$

where R_n is the cross rate of the transmission mode n, which can be estimated as

$$R_n = \sqrt{2\pi \frac{m\gamma_n}{\bar{\gamma}}} \frac{f_d}{\Gamma\left(m\right)} \left(\frac{m\gamma_n}{\bar{\gamma}}\right)^{m-1} \exp\left(-\frac{m\gamma_n}{\bar{\gamma}}\right) \tag{7}$$

where f_d denotes the Doppler frequency. Since the mobility of the MTs is not considered in this work, it is assumed that f_d is a fixed value (e.g., 10 Hz).

IV. Overhearing-Aided Redundancy Elimination

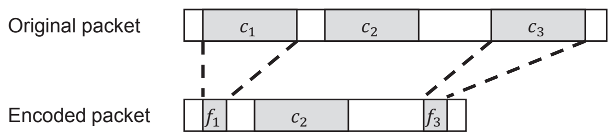
In OARE, the AP calculates the overhearing probabilities of redundant chunks in an incoming packet. After that, it computes a threshold for encoding chunks to minimize the expected packet transfer time. The following two subsections describe how to compute the overhearing probability and derive the optimal encoding threshold. After that, we will demonstrate the retransmission procedure and operations of AP and MT in OARE.

A. Computation of overhearing probability

In this subsection, we explain how to estimate the overhearing probability of the chunk in the cache table of the AP. Figure 4 shows an example scenario in which the current time is t_β and the AP has $packet_B$ destined to MT 1 in which the packet contains redundant chunk c_{Red}. The index of c_{Red} in the cache table is ω and it is assumed that c_{Red} was transmitted to MT 2 by belonging to $packet_A$ at time t_α. As a result, $V_{AP,\omega} = \{c_{AP,\omega}, f_{AP,\omega}, n_{AP,\omega}, t_{AP,\omega}\}$ was stored in the cache table of the AP where $c_{AP,\omega}$ and $t_{AP,\omega}$ are set to c_{Red} and t_α, respectively. $f_{AP,\omega}$ is computed from $c_{AP,\omega}$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

$$\mathrm{PER}\left(n,\nu,\bar{\gamma}\right)=\frac{1}{\pi\left(\nu,\bar{\gamma}\right)}\int_{\gamma_\nu}^{\gamma_{\nu+1}}\mathrm{PER}\left(n,\gamma\right)f_{\bar{\gamma}}\left(\gamma\right)d\gamma$$
$$=\begin{cases} 1, & \text{if } \gamma_{\nu+1}<\gamma_{pn} \\ \frac{1}{\pi(\nu,\bar{\gamma})}\frac{1}{\Gamma(m)}\left[\Gamma\left(m,\frac{m\gamma_\nu}{\bar{\gamma}}\right)-\Gamma\left(m,\frac{m\gamma_{pn}}{\bar{\gamma}}\right)+a_n\left(\frac{m}{\bar{\gamma}}\right)^m\frac{\Gamma(m,b_n\gamma_{pn})-\Gamma(m,b_n\gamma_{\nu+1})}{(b_n)^m}\right], & \text{if } \gamma_\nu\le\gamma_{pn}<\gamma_{\nu+1} \\ \frac{1}{\pi(\nu,\bar{\gamma})}\frac{a_n}{\Gamma(m)}\left(\frac{m}{\bar{\gamma}}\right)^m\frac{\Gamma(m,b_n,\gamma_{pn})-\Gamma(m,b_n\gamma_{\nu+1})}{(b_n)^m}, & \text{if } \gamma_{pn}<\gamma_\nu \end{cases} \tag{5}$$



$Q = \{c_1, c_2, c_3\}, R = \{\theta_1, \theta_2, \theta_3\}\ (\theta_3 > \theta_1 > \theta_2)$

$\delta = \theta_1, Q_\delta = \{c_1, c_3\}$

Fig. 5: An example scenario for encoding a packet. The original packet has three redundant chunks. Among them, two chunks are replaced with their corresponding fingerprints where their overhearing probabilities are greater than or equal to δ.

$n_{AP,\omega}$ indicates the transmission mode ID of $packet_A$ because $c_{AP,\omega}$ was a part of $packet_A$.

To estimate the overhearing probability of c_{Red} for MT 1 at t_β, the transmission mode of MT 1 at t_α should be first obtained since the overhearing was accomplished at that time. The transmission mode of MT d at the time slot t is denoted by $n_{d,t}$, $\forall n_{d,t} \in \{1, ..., N\}$. Then, the probability that MT d used the transmission mode i at t_α while it uses the transmission mode j at t_β can be expressed as

$$\mathrm{P}\left(n_{d,t_\alpha}=i\left|n_{d,t_\beta}=j\right.\right)=\frac{\mathrm{P}\left(n_{d,t_\beta}=j\left|n_{d,t_\alpha}=i\right.\right)}{\mathrm{P}\left(n_{d,t_\beta}=j\right)}\mathrm{P}\left(n_{d,t_\alpha}=i\right)=\frac{\pi\left(i,\bar{\gamma}\right)}{\pi\left(j,\bar{\gamma}\right)}\zeta_{i,j}^{(t_\beta-t_\alpha)} \tag{8}$$

where $\bar{\gamma}$ is the average received SNR of MT d and $\zeta_{i,j}^{(\Delta)}$ is the Δ-step transition probability of the FSMC model which means the state transition probability from state i to state j after Δ time slots. Consequently, the overhearing probability of c_{Red} for MT d can be estimated by using the corresponding cached item $V_{AP,\omega}$ and it is represented as (9) where $t_{IN,d}$ is the arrived time of MT d into the cell of the AP. It is obvious that MTs cannot receive any packets out of the cell, and therefore the overhearing probability is zero when $t_{AP,\omega} < t_{IN,d}$.

B. Derivation of δ

When a AP receives an incoming packet destined to an MT, the packet may have several redundant chunks as described in Figure 5. Let the packet have H redundant chunks denoted by $Q = \{c_1, c_2, ..., c_H\}$. For $c_i \in Q$, the overhearing probability for the destination MT is denoted by θ_i ($R = \{\theta_1, \theta_2, ..., \theta_H\}$). Since the chunks in the cache table were originally transmitted to different MTs via different packets at different times, their overhearing probabilities are totally different from each other. That is, some of them were successfully overheard by the destination MT, but some were not. Therefore, a set of chunks to be encoded, which have sufficiently high overhearing probabilities, should be carefully selected. In OARE, the optimal encoding threshold δ is dynamically computed for each incoming packet and only a set of chunks where the overhearing probabilities are greater than or equal to δ are encoded to minimize the expected packet transfer time.

TABLE II: Parameters for derivation of δ

Parameter	Definition
Q	Set of chunks of a packet
R	Set of overhearing probabilities of Q
δ	Encoding threshold
Q_δ	Subset of Q which are the corresponding chunks to R_δ
R_δ	Subset of R where all elements are greater than or equal to δ
L_C	Transmission delay of a chunk
L_F	Transmission delay of a fingerprint
L_N	Transfer time of an NAK message

Table II summarizes the parameters for the derivation of δ. Q denotes a set of chunks in an incoming packet and R represents their overhearing probabilities computed by Eq. (9). δ is the encoding threshold for the packet, which is chosen from R (i.e., $\delta \in R$). Then, R_δ and Q_δ consist of overhearing probabilities which are greater than or equal to δ and their corresponding chunks ($R_\delta \subset R$ and $Q_\delta \subset Q$), respectively. L_C denotes the transmission delay of a chunk (i.e., the amount of time required to push all of the chunk's bits into the wireless medium). Similarly, L_F denotes the transmission delay of a fingerprint. On the other hand, L_N represents the transfer time of a no-acknowledgement (NAK) message, which includes both the propagation delay and the transmission delay.

The problem of minimizing the expected packet transfer time can be converted to maximizing the benefit of encoded packet transmission. If a chunk is encoded by replacing the corresponding fingerprint, the packet length is reduced since the size of the fingerprint is much smaller than that of the chunk. Therefore, the transmission delay of the packet can be reduced as much as $L_C - L_F$ by encoding a chunk where L_F is smaller than L_C. However, after receiving the encoded packet, the destination MT may not decode the packet due to the absence of the fingerprints in its cache. In this case, the MT sends an NAK message for triggering the retransmission for the non-decodable chunks. After that, the AP sends the original chunks to the MT. Therefore, in case of decoding failure, L_N and L_C for each non-decodable chunk are additionally

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

$$\theta_{c_{Red},d}\left(V_{AP,\omega},t_\beta\right)=\begin{cases}\sum\limits_{i=1}^{N}\frac{\pi(i,\bar{\gamma})}{\pi\left(n_{d,t_\beta},\bar{\gamma}\right)}\zeta_{i,n_{d,t_\beta}}^{(t_\beta-t_{AP,\omega})}\left(1-\mathrm{PER}\left(n_{AP,\omega},i,\bar{\gamma}\right)\right), & \text{if } t_{AP,\omega}\geq t_{IN,d}\\ 0, & \text{if } t_{AP,\omega}<t_{IN,d}\end{cases} \tag{9}$$

$$E\left[B\right]=\begin{cases}\sum\limits_{\theta_j\in R_\delta}\left(\left(L_C-L_F\right)\theta_j-L_F\left(1-\theta_j\right)\right)-L_N\left(1-\prod\limits_{\theta_j\in R_\delta}\theta_j\right), & \text{if } R_\delta\neq\emptyset\\ 0, & \text{if } R_\delta=\emptyset\end{cases} \tag{10}$$

consumed to complete the packet delivery. Consequently, the expected benefit of the encoded packet transmission, $E\left[B\right]$, is represented as (10). If R_δ is an empty set, no chunk is encoded and the expected gain is simply zero. According to Eq. (10), the AP chooses δ to maximize the benefit $E[B]$, i.e., the AP conducts

$$\delta=\arg\max_{\theta\in R}E\left[B\right]. \tag{11}$$

To this end, expected benefits for all values in R_δ are computed by Eq. (10). For this procedure, different overhearing probabilities of the redundant chunks are multiplied to calculate the expected cost of delivering an NAK message (i.e., $L_N\left(1-\prod_{\theta_j\in R_\delta}\theta_j\right)$). When the number of redundant chunks of the packet is H, the corresponding computational complexity can be expressed as $O(H^2)$. After that, the maximum gain is found out among the set of calculated benefits by means of a simple selection algorithm of which the complexity is $O(H)$. As a result, the overall computational complexity for deriving δ can be represented by $O(H^2)$. In addition, the number of redundant chunks H has an upper bound since the length of a packet is constant (e.g., 1500 bytes) and thus, the optimal δ can be obtained in fixed time.

Finally, a set of chunks in which their overhearing probabilities are greater than or equal to δ are replaced by their fingerprints and the encoded packet is transmitted instead of the original packet. As shown in Figure 5, the optimal encoding threshold δ is computed as θ_1 and two chunks (c_1 and c_3) have sufficient overhearing probabilities that are greater than or equal to δ (i.e., $\theta_1\geq\delta$ and $\theta_3\geq\delta$). Consequently, these two chunks are replaced with the corresponding fingerprints (f_1 and f_3) and the encoded packet is transmitted to the MT instead of the original packet.

C. Chunk retransmission

When the AP transmits an encoded packet, it stores the original packet in its temporal buffer to cope with decoding failure events. If the MT could not overhear the encoded chunks in the past, those chunks do not exist in the cache table and the MT cannot decode the encoded packet. For such a decoding failure case, the MT requests the original chunks to the AP by transmitting an NAK message that contains a set of offsets of original chunks to be retransmitted. After receiving the NAK message, the AP transmits the original chunks to the MT.

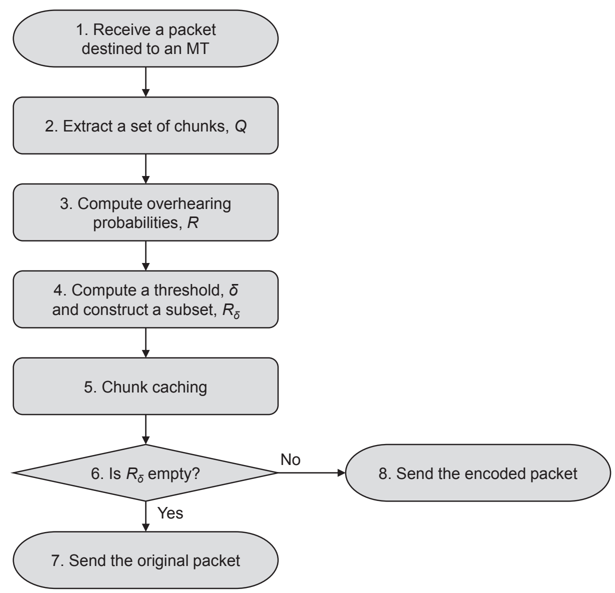


Fig. 6: AP Operation.

D. AP operation

Figure 6 describes the operation of a AP in OARE. Before the AP sends an incoming packet to the destination MT, the AP investigates redundant chunks of the incoming packet and encodes identified redundant chunks depending on their overhearing probabilities. By transmitting a small encoded packet rather than the original one, the transmission delay can be significantly reduced. The detailed operation of the AP can be summarized as follows.

- Step 1: First, the AP receives an incoming packet destined to an MT.
- Step 2: The AP extracts a set of chunks Q from the incoming packet and their fingerprints are computed by a hash function (e.g., SHA-1).
- Step 3: The AP computes the overhearing probability of each chunk in Q for the destination MT according to (9) and constructs a set of overhearing probabilities R for the destination MT.
- Step 4: The AP determines the optimal encoding threshold δ to minimize the expected packet transfer time from (11). Also, it constructs a set of overhearing probabilities R_δ in which all of elements are greater than or equal to δ.
- Step 5: The AP updates its cache table with the chunks that have overhearing probabilities less than δ. This is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

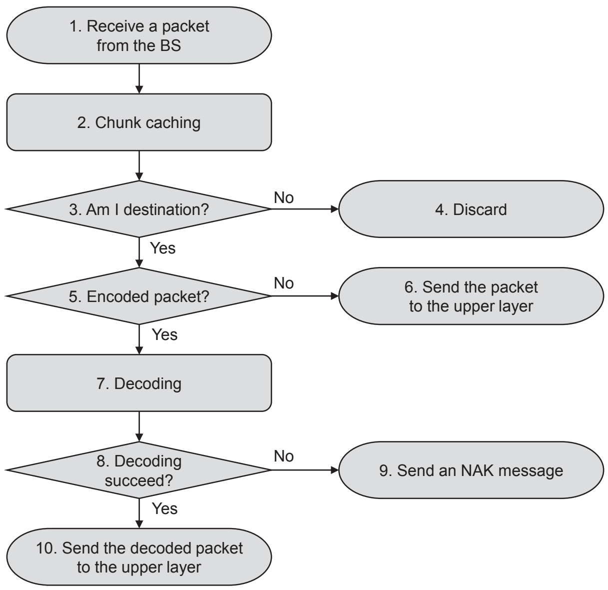


Fig. 7: MT operation.

because they will not be encoded (i.e., the original chunks are transmitted belonging to the transmitted packet) and thus other MTs can interpret only those chunks after overhearing.

- Steps 6–7: If R_δ is empty, the AP transmits the original packet to the destination MT since there is no gain to encode the packet.
- Step 8: If R_δ is not empty at Step 6, the AP encodes the packet depending on δ and transmits the encoded packet to the MT. In addition, the AP stores the original packet in its temporal buffer to cope with decoding failure events.

E. MT operation

Figure 7 describes the operation of an MT in OARE. The MT receives a packet which is intended to transmit to itself or not (i.e., overhear). After receiving the packet, the MT updates its cache table and checks the destination of the packet. Then the MT decodes the encoded packet which is destined to itself. If the decoding process is failed, the MT requests the original chunks to the AP. The detailed operation of the MT can be summarized as follows.

- Step 1: The MT receives a packet from the AP where the packet could be transmitted to the MT or obtained by overhearing.
- Step 2: Regardless of the actual destination of the packet, the MT updates its cache table with the chunks of the packet. Note that the encoded part (i.e., replaced by fingerprints) of the packet is excluded in this step. In contrast with the AP, the MT does not record the transmission mode and the current time slot.
- Steps 3–4: If the MT is not the destination of the packet, the packet is discarded.
- Steps 5–6: If the packet is not encoded, it is intactly sent to the upper layer.
- Step 7: If the received packet is encoded at Step 5, the MT attempts to reconstruct (i.e., decode) the packet. That is, the fingerprints of the encoded packet are replaced with the corresponding chunks in the cache table of the MT.
- Steps 8–9: If the decoding process is failed, the MT requests the original chunks by sending an NAK message to the AP. After receiving original chunks, the MT reconstructs the packet and sends it to the upper layer.
- Step 10: If the MT decodes the packet successfully at Step 8, the decoded one is sent to the upper layer.

V. Simulation Results

For performance evaluation, we develop a simulator using Java and conduct extensive simulations. In the simulation, 16000 files whose average size is about 250 Kbytes are used. In terms of network topology, one AP and five MTs are assumed. Each MT has a different average received SNR, $\bar{\gamma}$, to represent its own channel conditions. Specifically, the average received SNRs of five MTs are set to 4.2, 7.2, 11.1, 13.6, and 19.2 dB, respectively. To consider different popularities among the set of arbitrary files, a Zipf-like distribution is assumed which has been proved to describe the query pattern of web data access [10]. Accordingly, the relative probability of the ith most popular file is proportional to $1/i^\alpha$ where α determines the skewness in the Zipf-like distribution. For instance, for α = 1, the access probability of a file is strictly proportional to its popularity ranking (i.e., Zipf's law). On the other hand, for α = 0, the access probabilities for all files are the same. Let σ_i be the ith most popular file ($1 \le i \le$ Z). Then, the probability of accessing, σ_i, is given by

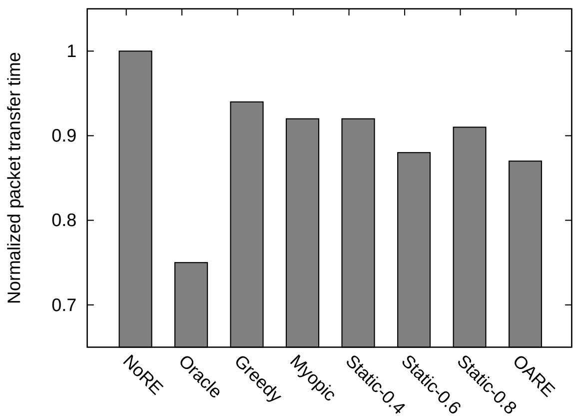
$$\sigma_i = \frac{\Omega}{i^\alpha} \tag{12}$$

where $\Omega = \left(\sum_{i=1}^{Z} (1/i^\alpha)\right)^{-1}$ and Z is the total number of files.

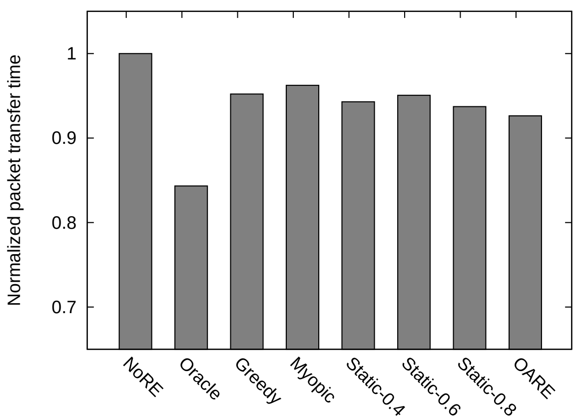
To verify the performance of OARE, we compare the packet transfer time of OARE with those of the following schemes.

- NoRE: This scheme does not exploit RE. All of packets are transmitted without any encoding procedure.
- Oracle: In this scheme, we assume that the AP knows the accurate cache information of every MT. Therefore, the AP can encode packets without any failure events based on the destination MT's cache table and achieve the best performance. The results of Oracle can be interpreted as the performance upper bound of RE.
- Greedy: To encode a set of chunks in an incoming packet, the AP only considers whether each chunk is stored or not in its cache table. If some chunks exist, the AP encodes those chunks without any estimation on that the chunks are also stored in the destination MT's cache. In other words, the AP assumes that every MT could overhear (or receive) those chunks in the past transmissions.
- Myopic: In this scheme, the AP assumes that wireless channel conditions are not changed and thus it simply compares the current transmission mode of the destination MT, $n_{d,\tau}$, and the past transmission mode of the redundant chunk, n_ω, to determine whether the chunk was overheard or not by the destination MT. If $n_{d,\tau}$ is greater than or equal to n_ω, the AP regards that the destination

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology



(a) Scenario 1 ($\alpha = 0.8$, $S = 2^{19}$, and $m = 0.5$).



(b) Scenario 2 ($\alpha = 0.7$, $S = 2^{18}$, and $m = 1.0$).

Fig. 8: Packet transfer time for 1 Gbytes traffic in two different scenarios. All of values are normalized to NoRE.

MT had sufficiently good channel conditions to overhear a chunk, and therefore it encodes the chunk.

- Static-p: To determine which of chunks will be encoded, the AP calculates the overhearing probabilities of redundant chunks according to (9). Then, it encodes the chunks where their overhearing probabilities are greater than or equal to a predefined fixed threshold, p. In other words, the encoding threshold is fixed at p.
- OARE: In OARE, the AP computes the optimal encoding threshold δ for each packet to minimize the expected packet transfer time. After calculating the overhearing probabilities of redundant chunks, the AP computes δ by (11) and encodes a set of chunks where the overhearing probabilities are greater than or equal to δ.

A. *Performance comparison*

Figure 8(a) and 8(b) show the packet transfer times of the aforementioned schemes to transmit 1 Gbytes traffic in two different simulation configurations. All results are normalized by the packet transfer time of NoRE. In Scenario 1, the Zipf skewness parameter α, the cache table size S, and the Nakagami fading parameter m are configured as 0.8, 2^{19} items, and 0.5, respectively. Meanwhile, in Scenario 2, they are set to 0.7, 2^{18}, and 1.0, respectively. Obviously, Oracle shows the lowest transfer time (i.e., 0.75 in Figure 8(a) and 0.84 in Figure 8(b)) since it encodes chunks as much as possible without any decoding failures. On the other hand, Greedy and Myopic achieve only 0.94 and 0.92 (0.95 and 0.96), respectively, in Scenario 1 (Scenario 2) whereas the performance of Static-p is affected by p. In particular, Static-0.6 and Static-0.8 show comparable transfer times to OARE (or the best performance among static methods) in Scenario 1 and Scenario 2, respectively. To conclude, different values of p should be employed to attain the optimal performance in static methods, which is not a trivial task. On the contrary, it can be found that the dynamic approach of OARE provides the best performance except Oracle in both scenarios.

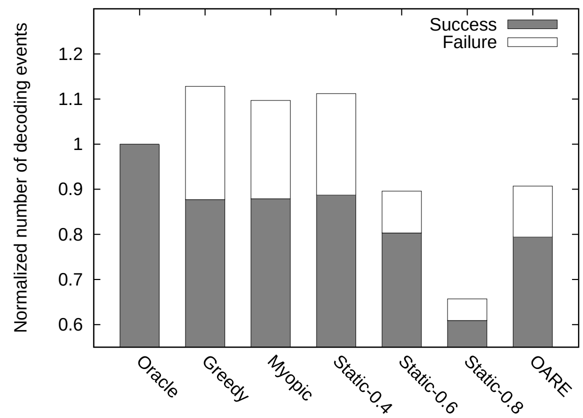


Fig. 9: Number of decoding success and failure events ($\alpha = 0.8$, $S = 2^{19}$, and $m = 0.5$). All of values are normalized to Oracle

To analyze the first result (i.e., Figure 8(a)), we compare the number of decoding success and failure events at the MTs during the simulations. In Figure 9, the gray and white bars indicate the number of decoding success events and the number of decoding failure events, respectively. The sum of gray and white bar is the total number of encoding events at the AP. All of values are normalized by the number of encoding events of Oracle, in which no decoding failure events occur since the encoding is performed based on the perfect cache information of MTs. It can be seen that the most number of chunks are encoded in Greedy (i.e., 1.128), however the most number of chunks cannot be decoded (i.e., 0.251). If the popularity of each request is not sufficiently skewed (i.e., α is low), Greedy can cause severe performance degradation. Although the number of decoding success events of Myopic is similar to that of Greedy, Myopic has the reduced number of decoding failure events and thus it can reduce the transmission time as shown in Figure 8(a). In Static-p, as p increases, not only the number of encoding events but also the number of decoding failure events decrease. That is, when p is large, higher accuracy of encoding can be obtained; however less opportunity to reduce the packet transfer time is provided. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

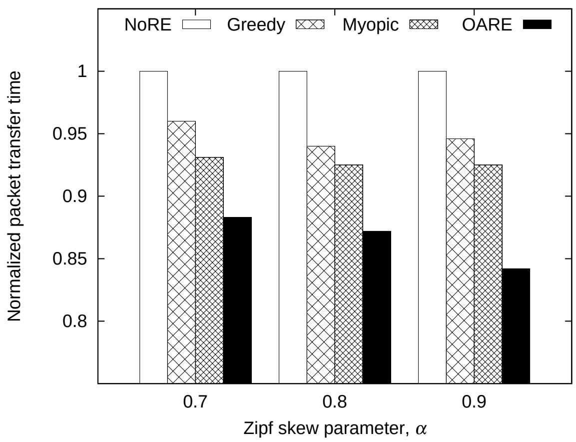


Fig. 10: Packet transfer time depending on Zipf skew parameter, α ($S = 2^{19}$ and $m = 0.5$). All of values are normalized to NoRE when α is 0.7.

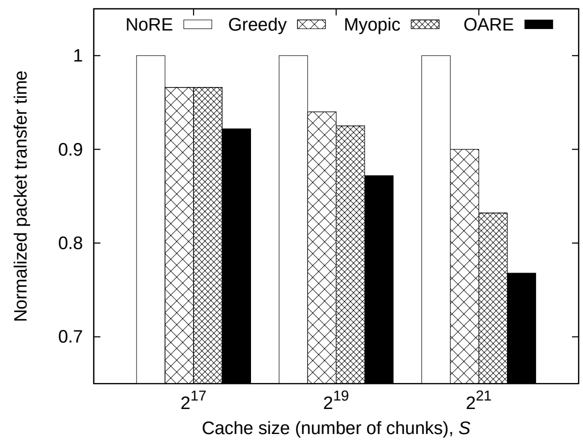


Fig. 11: Packet transfer time depending on the size of cache table, S ($\alpha = 0.8$ and $m = 0.5$). All of values are normalized to NoRE when S is 2^{17}.

numbers of decoding success and failure events in OARE are 0.794 and 0.113, respectively. In spite of that Static-0.6 shows better results (i.e., 0.803 and 0.093 for decoding success and failure events, respectively), the packet transfer time of OARE is still shorter than that of Static-0.6 as shown in Figure 8(a). This result reveals that the encoding threshold δ should be dynamically selected for each packet. If δ is configured by a static manner, the expected benefit of OARE cannot be maximized since it is dependent on not only the overhearing probabilities but also the number of encoded chunks according to Eq. (10). Note that additional delay for an NAK message is needed regardless of the number of non-decodable chunks.

B. Effect of skewness parameter α

Figure 10 illustrates the packet transfer time depending on the Zipf skew parameter, α. All of values are normalized by the packet transfer time of NoRE when α is 0.7. m and S are fixed as 0.5 and 2^{19} items, respectively. Note that α defines how much popularity of data is skewed. A larger α means that the popularity curve is more skewed, which implies that the top-ranked objects receive a higher fraction of the requests and the redundancy of network traffic is increased accordingly. From Figure 10, it can be seen that the performances of Greedy and Myopic are not improved evidently as α increases. Even worse, the transfer time of Greedy is slightly increased when α is changed from 0.8 to 0.9. This is because Greedy does not employ any accurate estimation technique for the overhearing probability. On the other hand, the packet transfer time of OARE is remarkably reduced with the increase of α. For example, when α is 0.7, the normalized transfer time is 0.883 and it is reduced to 0.872 and 0.842 when α is 0.8 and 0.9, respectively. To conclude, accurate estimation of the overhearing probability as in OARE is required to leverage higher redundancy.

C. Effect of cache size S

Figure 11 describes the packet transfer times depending on the cache size, S, when α and m are configured as 0.8 and 0.5,

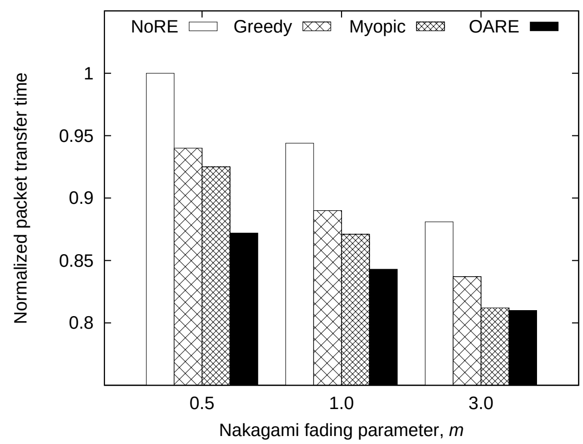


Fig. 12: Packet transfer time depending on Nakagami fading parameter, m ($\alpha = 0.8$ and $S = 2^{19}$). All of values are normalized to NoRE when m is 0.5.

respectively. In the proposed cache architecture, $log_2 S$ bits of a fingerprint are exploited to index the corresponding chunk in the cache table. Therefore, as S increases, the collision probability that two different chunks point the same index can be reduced. That is, replacement frequency of each chunk is decremented. In such a situation, the AP often refers the old items in the cache table, and therefore a valid inference of the channel condition when the referenced item (i.e, redundant chunk) is originally transmitted becomes more important. In OARE, by leveraging FSMC, the past channel condition can be accurately estimated. As a result, it can be seen that the transfer time of OARE is further reduced by increasing the cache size, i.e., the transfer time decreases from 0.922 to 0.768 when the cache size increases from 2^{17} to 2^{21}.

D. Effect of fading parameter m

Figure 12 shows the performance variations for different values of the Nakagami fading parameter, m, when α and S are configured as 0.8 and 2^{19} items, respectively. Each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

value is normalized to NoRE when m is 0.5. In the Nakagami fading channel model, m represents the channel stability. As m increases, wireless channel condition becomes better and more stable, and therefore the packet transfer time of NoRE is further reduced. Of course, the transfer times of other schemes decrease as m increases. It can be seen that OARE obtains the lowest transmission time in every scenario. Interestingly, the degree of degradation is the most sharp in Myopic. This is because as the condition of wireless channel becomes more stable, the state transitions occur infrequently and thus, a simple comparison between the current transmission mode and the past transmission mode can work well. For example, when m is 0.5 (i.e., wireless channel varies rapidly), the packet transfer time of Myopic is slightly lower than Greedy. Meanwhile, when m is 3.0, the results of Myopic and OARE are almost the same (i.e., 0.812 and 0.81, respectively).

VI. Conclusions

In this paper, we have proposed an overhearing-aided redundancy elimination (OARE) scheme by leveraging the broadcast nature of wireless medium in multi-rate wireless networks. We have developed novel analytical models to estimate an overhearing probability and compute an optimal encoding threshold value. Extensive simulation results demonstrate that OARE can reduce the packet transfer time by 7.8–23.2% compared with non-RE (or non-encoded) transmissions. It is envisioned OARE can be employed to improve the network throughput in WiFi systems in a cost-effective manner (i.e., without increasing the network bandwidth). In our future work, we will analyze more features such as user mobility and interference among multiple cells in order to deploy OARE in next-generation cellular networks. In addition, we will investigate how to take advantage of sociality among mobile users to further enhance the performance of OARE.

References

[1] Y. Shim, G. Park, I. Jang, and S. Pack, "Overhearing-Aided Redundancy Elimination in Multi-Rate Wireless Networks," in *Proc. AsiaFI 2013 Summer School*, Hongkong, China, August 2013.

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," white paper, Cisco Systems, February 2014.

[3] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile Data Offloading Through A Third-Party WiFi Access Point: An Operator's Perspective," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5340–5351, October 2014.

[4] N. Cheng, N. Lu, N. Zhang, X. Shen, and J. W. Mark, "Vehicular WiFi Offloading: Challenges and Solutions," *Elsevier Vehicular Communications*, vol. 1, no. 1, pp. 13–21, January 2014.

[5] S. I. Sou, "Mobile Data Offloading With Policy and Charging Control in 3GPP Core Network," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3481–3486, September 2013.

[6] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile Data Offloading: How Much Can WiFi Deliver?" *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 536–551, April 2013.

[7] G. Lee, I. Jang, S. Pack, and X. Shen, "FW-DAS: Fast Wireless Data Access Scheme in Mobile Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4260–4272, August 2014.

[8] 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 Communications Magazine*, vol. 52, no. 2, pp. 131–139, February 2014.

[9] W. Wu, J. Cao, and X. Fan, "Design and Performance Evaluation of Overhearing-Aided Data Caching in Wireless Ad Hoc Networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 3, pp. 450–463, March 2013.

[10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," in *Proc. IEEE INFOCOM 1999*, New York, USA, March 1999.

[11] S. Pack, H. Rutagemwa, X. Shen, J .W. Mark, and K. Park, "Proxy-based Wireless Data Access in Mobile Hotspots," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 5, pp. 3165–3177, September 2008.

[12] Y. Zhang and N. Ansari, "On Protocol-Independent Data Redundancy Elimination," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp 455–472, 1st Quarter 2014.

[13] N. Spring and D. Wetherall, "A Protocol Independent Technique for Eliminating Redundant Network Traffic," in *Proc. ACM SIGCOMM 2000*, Stockholm, Sweden, August 2000.

[14] A. Anand, C. Muthukrishnan, A. Akella, and R. Ramjee, "Redundancy in Network Traffic: Findings and Implications," in *Proc. ACM SIGMETRICS 2009*, Seattle, USA, June 2009.

[15] B. Aggarwal, A. Akella, A. Anand, A. Balachandran, P. Chitnis, C. Muthukrishnan, R. Ramjee, and G. Varghese, "EndRE: An End-System Redundancy Elimination Service for Enterprises," in *Proc. USENIX NSDI 2010*, San Jose, USA, April 2010.

[16] A. Anand, A. Gupta, A. Akella, S. Seshan, and S. Shenker, "Packet Caches on Routers: The Implications of Universal Redundant Traffic Elimination," in *Proc. ACM SIGCOMM 2008*, Seattle, USA, August 2008.

[17] A. Anand, V. Sekar, and A. Akella, "SmartRE: An Architecture for Coordinated Network-wide Redundancy Elimination," in *Proc. ACM SIGCOMM 2009*, Barcelona, Spain, August 2009.

[18] E. Zohar, I. Cidon, and O. Mokryn, "PACK: Prediction-Based Cloud Bandwidth and Cost Reduction System," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 39–51, February 2014.

[19] L. Yu, K. Sapra, H. Shen, and L. Ye, "Cooperative End-to-End Traffic Redundancy Elimination for Reducing Cloud Bandwidth Cost," in *Proc. IEEE ICNP 2012*, Austin, USA, October 2012.

[20] Z. Zhuang and R. Sivakumar, "Wireless Memory: Eliminating Communication Redundancy in Wi-Fi Networks," in *Proc. IEEE WoWMoM 2011*, Lucca, Italy, June 2011.

[21] S. Shen, A. Gember, A. Anand, and A. Akella, "REfactor-ing Content Overhearing to Improve Wireless Performance," in *Proc. ACM MobiCom 2011*, Las Vegas, USA, September 2011.

[22] C. Lumezanu, K. Guo, N. Spring, and B. Bhattacharjee, "The Effect of Packet Loss on Redundancy Elimination in Cellular Wireless Networks," in *Proc. ACM IMC 2010*, Melbourne, Australia, November 2010.

[23] E. Zohar, I. Cidon, and O. O. Mokryn, "Celleration: Loss-Resilient Traffic Redundancy Elimination for Cellular Data," in *Proc. ACM HotMobile 2012*, San Diego, USA, February 2012.

[24] F. Sun, T. M. Kim, A. J. Paulraj, E. de Carvalho, and P. Popovski, "Cell-Edge Multi-User Relaying with Overhearing," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1160–1163, June 2013.

[25] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-Layer Combining of Adaptive Modulation and Coding With Truncated ARQ Over Wireless Links," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 17461755, September 2004.

[26] G. L. Stüber, *Principles of Mobile Communication*, 2nd Edition, Norwell, Massachusetts: Kluwer, 2001.

[27] "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band," IEEE Standard 802.11a-1999-Part 11, 1999.

[28] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing With Adaptive Modulation and Coding Over Wireless Links: Cross-Layer Analysis and Design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1142–1153, May 2005.

[29] H. S. Wang and N. Moayeri, "Finite-State Markov Channel - A Useful Model for Radio Communication Channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, February 1995.

Gwangwoo Park received the B.S. and M.S. degrees from Korea University, Seoul, Korea, in 2010 and 2012, respectively. He is currently working towards his Ph.D. degree in Depart of IT Convergence, Korea University, Seoul, Korea. His research interests include content networking, multimedia streaming, information-centric networking, and Future Internet.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2438822, IEEE Transactions on Vehicular Technology

Youngjun Shim received the B.S. degree from Kwangwoon University, Seoul, Korea, in 2012 and the M.S. degree from Korea University, Seoul, Korea, in 2014. He is currently working at Samsung Electronics. His research interests include content networking, multimedia networking, and Future Internet.

Insun Jang received the B.S. degree from Korea University, Seoul, Korea, in 2011. He is currently an M.S. and Ph.D. integrated course student in School of Electrical Engineering, Korea University, Seoul, Korea. His research interests include content delivery networks, multimedia networking, software defined networking, and Future Internet.

Sangheon Pack (S'02–M'05–SM'11) received the B.S. (magna cum laude) and Ph.D. degrees from Seoul National University, Seoul, Korea, in 2000 and 2005, respectively, both in computer engineering. In 2007, he joined the faculty of Korea University, Seoul, Korea, where he is currently an Associate Professor in the School of Electrical Engineering. From 2005 to 2006, he was a Postdoctoral Fellow with the Broadband Communications Research Group, University of Waterloo, Waterloo, ON, Canada. He was the recipient of KICS (Korean Institute of Communications and Information Sciences) Haedong Young Scholar Award 2013, IEEE ComSoc APB Outstanding Young Researcher Award in 2009, LG Yonam Foundation Overseas Research Professor Program in 2012, and Student Travel Grant Award at the IFIP Personal Wireless Conference (PWC) 2003. From 2002 to 2005, he was a recipient of the Korea Foundation for Advanced Studies Computer Science and Information Technology Scholarship. From 2010 to 2012, he was a head of Department of Samsung IT Convergence in Korea University. In 2003, he was a Visiting Researcher at Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin, Germany. He was a publication co-chair of IEEE INFOCOM 2014/ACM Mobihoc 2015, a co-chair of IEEE VTC 2010-Fall transportation track, a co-chair of IEEE WCSP 2013 wireless networking symposium, a TPC vice-chair of ICOIN 2013, and a publicity co-chair of IEEE SECON 2012. He is an editor of Journal of Communications Networks (JCN), IET Communications, and International Journal of Distributed Sensor Networks. He is a senior member of the IEEE. His research interests include Future Internet, SDN/ICN/DTN, mobility management, mobile cloud networking, multimedia networking, and vehicular networks.

Xuemin (Sherman) Shen received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. Dr. Shens research focuses on resource management in interconnected wireless/wired networks, wireless network security, wireless body area networks, vehicular ad hoc and sensor networks. Dr. Shen serves/served as the Editor-in-Chief for IEEE Network, Peer-to-Peer Networking and Application, and IET Communications; a Founding Area Editor for IEEE Transactions on Wireless Communications; an Associate Editor for IEEE Transactions on Vehicular Technology, Computer Networks, and ACM/Wireless Networks, etc.; and the Guest Editor for IEEE JSAC, IEEE Wireless Communications, IEEE Communications Magazine, and ACM Mobile Networks and Applications, etc. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.