PAPER Special Section on Information Theory and Its Applications

Coded Caching in Multi-Rate Wireless Networks*

Makoto TAKITA[†a)], **Masanori HIROTOMO**[††], ***Members***, ***and*** **Masakatu MORII**[†††], ***Fellow***

SUMMARY The network load is increasing due to the spread of content distribution services. Caching is recognized as a technique to reduce the peak network load by storing popular content into memories of users. Coded caching is a new caching approach based on a carefully designed content placement to create coded multicasting opportunities. Coded caching schemes in single-rate networks are evaluated by the tradeoff between the size of memory and that of delivered data. For considering the network with multiple transmission rates, it is crucial how to operate multicast. In multicast delivery, a sender must communicate to intended receivers at a rate that is available to all receivers. Multicast scheduling method of determining rates to deliver are evaluated by throughput and delay in multi-rate wireless networks. In this paper, we discuss coded caching in the multi-rate wireless networks. We newly define a measure for evaluating the coded caching scheme as coded caching delay and propose a new coded caching scheme. Also, we compare the proposed coded caching scheme with conventional coded caching schemes and show that the proposed scheme is suitable for multi-rate wireless networks.

key words: *coded caching, multi-rate network, delay, wireless networks, scheduling method*

1. Introduction

The peak network load is getting larger due to the demand for video streaming services using the Internet. Caching is recognized as one solution to reduce a peak network load [1]. When the network load is low, popular content is stored in the cache memories of end-users or the mass memories of cache servers based on a caching strategy. If the end-users can recover their requests by using the content stored in their memories, the network load can be reduced.

Coded caching is a new caching approach based on a carefully designed content placement to create coded multicasting opportunities [2], [3]. It can be achieved by making coded messages based on the content not only in the memory of each user but also in the memories of all users among the network. The placement phase is carefully designed to provide coded messages for multiple users. The transmission of the coded messages can use a network resource effectively by utilizing the multicast. There are various studies for the coded caching schemes with extensive network models. Network models that each user has a non-uniform demand are discussed in [4]–[6]. Network models with various cache sizes are discussed in [7] and [8]. Also, the models with distinct file sizes are discussed in [9]. There are some schemes in which not only one kind of heterogeneity is assumed, but also a plurality of kinds of heterogeneity is considered. For example, a model that simultaneously considers uneven file size, memory size, and file popularity is discussed in [10]. Also, a model where cache size and link quality are different for two user cases is discussed in [11]. These studies are evaluated by the tradeoff between the size of memory and the size of delivered data. When considering channel quality, e.g., the non-uniform transmission rate of the channel, we consider that it is necessary to be evaluated not only by the above tradeoff but also by a latency. In this paper, we discuss what problems are involved in the coded caching schemes for non-uniform transmission rates.

The multicast is known as an efficient method for delivering the same data to multiple receivers. However, the multicast is not always the most efficient method to transmit to all intended receivers in the multi-rate wireless networks where multiple receivers communicate with different transmission rates. The transmitter determines a transmission rate depending on the network environments in the wireless network. For example, the transmission rate is large for the receiver, which is near the access point, and the transmission rate is low for the receiver, which is distant from the access point. For considering multicasting to these receivers, the transmitter needs to transmit with the small transmission rate for the distant receiver. It leads to a reduction in the throughput of the entire network. Scheduling methods for selecting the transmission rate are essential to avoid this problem. The scheduling that uses the transmission delay as a measure of evaluation is discussed in [12]. The tradeoff relationship between the throughput and the delay is discussed in [13].

In this paper, we discuss a coded caching problem in the multi-rate wireless networks. We are not concerned here with the content popularity. The purpose of this paper is to clarify the influence of the multi-rate environment on the coded caching schemes. We define a coded caching delay as the measure to evaluate the coded caching schemes in the multi-rate wireless networks. We point out the weakness of the conventional coded caching scheme for the network with the different sizes of memory, called the zero-padding

Manuscript received January 25, 2020.
Manuscript revised June 1, 2020.

[†]The author is with the School of Social Information Science, University of Hyogo, Kobe-shi, 651-2197 Japan.
[††]The author is with the Graduate School of Science and Engineering, Saga University, Saga-shi, 840-8502 Japan.
[†††]The author is with the Graduate School of Engineering, Kobe University, Kobe-shi, 657-8501 Japan.

*The material in this paper was presented in part at 2018 IEEE International Conference on Communications Workshops, ICC Workshops 2018 [17].

a) E-mail: takita@sis.u-hyogo.ac.jp
DOI: 10.1587/transfun.2020TAP0013

scheme [7], when the delay is used as a measure to evaluate. For this problem, we consider reducing the delay by devising a delivery algorithm and propose a new coded caching scheme in the multi-rate wireless networks. Also, we compare the proposed scheme with the uncoded caching scheme and the zero-padding scheme.

The remainder of the paper is organized as follows. Section 2 describes the model of multi-rate wireless networks considered in this paper. Section 3 reviews coded caching schemes in single-rate networks and describes measures to evaluate scheduling methods in multi-rate networks. Section 4 defines a measure to evaluate the coded caching schemes as in the multi-rate network. Also, it presents a weakness of a conventional scheme in delay and proposes a new coded caching scheme to conquer the weak point. Section 5 compares the proposed coded caching scheme with conventional coded caching schemes and shows that the proposed scheme is suitable for multi-rate wireless networks.

2. Network Model

In this section, we describe a network model discussed in this paper.

We consider the multi-rate wireless network model illustrated in Fig. 1. The network consists of one sender and K receivers. The sender s has a database of N files $w_1, w_2, \ldots, w_N$ of each size F bits. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the index set of the files, and let $\mathcal{K} = \{1, 2, \ldots, K\}$ be the index set of the receivers. The receiver $i \in \mathcal{K}$ has a memory of size $M_i F$ bits. The quantity $M_i \in [0, N)$ is the normalized size of the memory by F and assume $M_1 \leq M_2 \leq \cdots \leq M_K$ without loss of generality. Let $C = \{C_1, C_2, \ldots, C_L\}$ be the set of the transmission rates available to the sender and assume $C_1 \leq C_2 \leq \cdots \leq C_L$. Let $c_k \in C$ be the maximum transmission rate which can be used between the sender s and the receiver k. The receiver k can use rate c_k or less in the set C.

The caching system operates in two phases, which are a placement phase and a delivery phase. In the placement phase, the receivers store contents related to the N files in their memories without any prior knowledge of future requests. In the delivery phase, each receiver requests one of the N files in the database, and the sender makes transmission messages and sends them to the receivers by multicasting. Let $d_k \in \mathcal{N}$ be the request of the receiver k, and it denotes the index of the requested file. Let $d = (d_1, d_2, \ldots, d_K)$ be the vector of requests. The sender makes a message of size $R^d F$ bits based on the request vector d, where R^d is the normalized by F, and sends it to receivers.

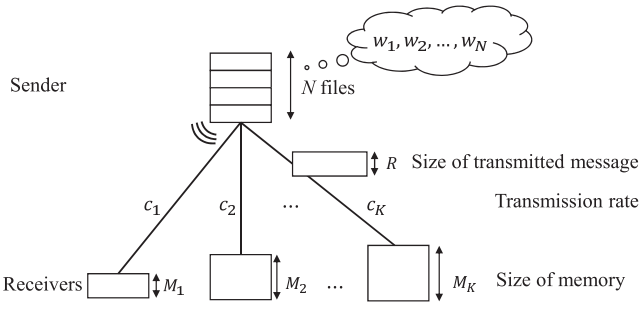


Fig. 1 Network model.

The limitations and issues to discuss for each phase are as follows: The placement phase operates when the network load is small, so the limitation of the network load is not considered. The problem in this phase is what content is stored in memories of limited size. The delivery phase operates when the network load is high, so the limitation of the network load is considered The problem in this phase is minimizing the size of delivering messages. At the time, the content stored in the memories can be used.

In this paper, we propose a coded caching scheme suitable for this model. The scheme is compared with the existing method based on the size of the message and the transmission delay. In order to carry out a discussion that does not depend on the file size, we used M_i and R^d which are the size normalized by file size for the following sections.

3. Related Works

In this section, we first review an existing approach to the coded caching problem in networks with non-uniform cache sizes. Besides, we will describe scheduling for determining a transmission rate to be used when performing multicast delivery in a multi-rate environment.

3.1 Coded Caching

The network load is increasing due to the expansion of content distribution services. The load tends to vary with time. For example, the load is small in the middle of the night, and the load is large in the evening. Caching is recognized as one solution to reduce the network load during peak times by prefetching popular content into memories of receivers. Coded caching [2], [3] is a new caching approach based on a carefully designed content placement to create coded multicasting opportunities.

The coded caching problem in the networks with a non-uniform size of memories of receivers is discussed in [7]. They assume that the receivers have memories of different size $M_1, M_2, \ldots, M_K$, but do not take into account the differences in the transmission rates, i.e., $c_1 = c_2 = \cdots = c_K$. Algorithm 1 is the coded caching scheme for this setting. $\mathcal{U}\backslash\{l\}$ denotes to exclude receiver l from the subset of receivers $\mathcal{U}$. For a receiver $l \in \mathcal{U}$, $w_{d_l,\mathcal{U}\backslash\{l\}}$ represents a partial file that is stored in the cache of receivers in $\mathcal{U}\backslash\{l\}$ but is not stored in the cache of other receivers in $\mathcal{K}$. $x_{\mathcal{U}}$ represents a coded file that is sent to receivers in $\mathcal{U}$. For simplicity, we describe $w_{d_l,\{1,2,3\}}$ as $w_{d_l,123}$ in this paper. $|w|$ is the size of the file w. The operator $\|$ means a concatenation of bit strings. When encoding data of different sizes, this scheme encodes by XORing data after zero-padding to scale to the maximum size.

Algorithm 1 Decentralized coded caching scheme with non-uniform cache size

Placement Phase
for $(k = 0; k < K; k++)$ **do**
  for $(n = 0; n < N; n++)$ **do**
    receiver k randomly prefetches $M_k F/N$ bits of content n;
  end for
end for
Delivery Phase
for $(k = K; k > 0; k--)$ **do**
  for each subset $\mathcal{U}$ of k receivers **do**
    Maxsize $\leftarrow \max_{i\in\mathcal{U}} |w_{d_i,\mathcal{U}\setminus\{i\}}|$;
    $x_{\mathcal{U}} \leftarrow$ all-zero vector of length Maxsize bits;
    for $l \in \mathcal{U}$ **do**
      if $|w_{d_l,\mathcal{U}\setminus\{l\}}| <$ Maxsize **then**
        temp $\leftarrow$ all-zero vector of length (Maxsize $- |w_{d_l,\mathcal{U}\setminus\{l\}}|$) bits;
        $w_{d_l,\mathcal{U}\setminus\{l\}} \leftarrow w_{d_l,\mathcal{U}\setminus\{l\}} \| \text{temp}$;
      end if
      $x_{\mathcal{U}} \leftarrow x_{\mathcal{U}} \oplus w_{d_l,\mathcal{U}\setminus\{l\}}$;
    end for
    Multicast the coded data $x_{\mathcal{U}}$ to receivers in $\mathcal{U}$.
  end for
end for

Table 1 The messages transmitted by the zero-padding scheme in Example 1.

Coded transmitted message	Size
$x_{123} = w_{d_1,23} \overline{\oplus} w_{d_2,13} \overline{\oplus} w_{d_3,12}$	$\left(1-\frac{M_1}{N}\right)\frac{M_2}{N}\frac{M_3}{N}$
$x_{12} = w_{d_1,2} \overline{\oplus} w_{d_2,1}$	$\left(1-\frac{M_1}{N}\right)\frac{M_2}{N}\left(1-\frac{M_3}{N}\right)$
$x_{13} = w_{d_1,3} \overline{\oplus} w_{d_3,1}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\frac{M_3}{N}$
$x_{23} = w_{d_2,3} \overline{\oplus} w_{d_3,2}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\frac{M_3}{N}$
$x_1 = w_{d_1,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_2 = w_{d_2,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_3 = w_{d_3,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$

Example 1 Suppose that a network consists of the sender s that has 3 files $\{w_1, w_2, w_3\}$ and 3 receivers $\{u_1, u_2, u_3\}$, i.e., $N = 3$ and $K = 3$. The receiver u_i, $i \in \{1, 2, 3\}$ has the memory of size M_i and vector of requests is $d = (d_1, d_2, d_3)$. Table 1 shows the messages transmitted based on Algorithm 1. In this table, the operator $\overline{\oplus}$ refers to the bitwise XOR operation after the zero-padding for scaling to the largest file.

The message x_{123} is made by encoding subfiles $w_{d_1,23}$, $w_{d_2,13}$, and $w_{d_3,12}$ that are different size and is sent to the receivers $\{u_1, u_2, u_3\}$. The expected size of each subfile is given by:

$$|w_{d_1,23}| = \left(1-\frac{M_1}{N}\right)\frac{M_2}{N}\frac{M_3}{N}, \quad (1)$$

$$|w_{d_2,13}| = \frac{M_1}{N}\left(1-\frac{M_2}{N}\right)\frac{M_3}{N}, \quad (2)$$

$$|w_{d_3,12}| = \frac{M_1}{N}\frac{M_2}{N}\left(1-\frac{M_3}{N}\right), \quad (3)$$

where $|w_{d_1,23}| \geq |w_{d_2,13}| \geq |w_{d_3,12}|$ from $M_1 \leq M_2 \leq M_3$. To XOR these files, the small files are needed to scale them to the largest file $w_{d_1,23}$ using zero-padding, as shown in Fig. 2. ■

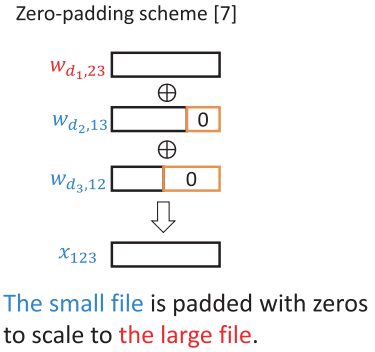


Fig. 2 An example of encoding using the zero-padding scheme.

Other studies considering heterogeneous file sizes are also based on the above zero-padding solution [10], [11]. In this paper, we point out that wasteful files are distributed when using a zero-padding scheme in a multi-rate environment, and propose a new delivery scheme.

3.2 Multi-Rate Wireless Networks

In wireless networks, a sender determines data transmission rates depending on physical distances between the sender and receivers, and the presence or absence of obstacles, and the like. For example, in IEEE 802.11b, which is one of the wireless LAN standards, 1 Mbps, 2 Mbps, 5.5 Mbps, and 11 Mbps are used according to conditions.

The multicast is well known as an efficient method for delivering the same data to multiple receivers. However, it is necessary to use the smallest transmission rate among the receivers that can communicate at different transmission rates in the multi-rate wireless networks. This problem causes a decrease in the communication efficiency of the whole network. Scheduling is important to avoid this problem. It decides the transmission rates which the receivers will use.

As a standard measure of communication in the network, there are throughput and a transmission delay. The throughput represents the amount of information that is processed per unit time. The transmission delay represents the time until the receiver receives whole the messages. For multicast in the multi-rate wireless networks, the scheduling with the minimum delay [12] and the tradeoff relationship between the throughput and the delay [13] are discussed.

We introduce the multicast throughput and the transmission delay defined by [13]. At first, the set of all transmission possibilities is considered. We assume that a sender s transmits a message x to K receivers. Let $\mathcal{T}_x = \{t_1, t_2, \ldots, t_m\}$ be the set of all scheduling that the sender s can consider. Each schedule is defined by the set of transmission rates to be used. For example, $t = \{(u_1, u_2)_{c_a}, (u_3)_{c_b}\}$ means that the receivers u_1 and u_2 receive with the transmission rate c_a and the receiver u_3 receives with the transmission rate c_b. For the message x and the scheduling t, $L_{x,t}$ is the number of carried-out transmissions and $C_{x,t}$ is the set of used transmission rates. $L_{x,t}$ is equal to the number of subsets of receivers in the scheduling t.

For the set of scheduling, a multi-rate multicast throughput and a multi-rate multicast delay are defined as follows.

Definition 1 (Multi-rate multicast throughput) [13]. The multicast throughput is the average rate of packets successfully delivered to multicast receivers. It is denoted by (4) when the sender s sends a message x with a scheduling t.

$$\mathrm{TH}_{x,t} \triangleq \frac{\sum_{c \in C_{x,t}} n_c c}{L_{x,t}}, \tag{4}$$

where n_c is the number of receivers that are received with the transmission rate c.

The multi-rate multicast throughput is the multicast throughput of the best scheduling. It is defined by

$$\mathrm{TH}_{x,\max} \triangleq \max_{t \in \mathcal{T}_x} \mathrm{TH}_{x,t}. \tag{5}$$

Definition 2 (Multi-rate multicast delay) [13]. The multicast delay is the total time spent to communicate a multicast packet to all receivers. It is denoted by (6) when the sender s sends a message x with a scheduling t.

$$\mathrm{TT}_{x,t} = \sum_{c \in C_{x,t}} \frac{|x|}{c}. \tag{6}$$

The multi-rate multicast delay is the multicast delay of the best scheduling. It is defined by

$$\mathrm{TT}_{x,\min} \triangleq \min_{t \in \mathcal{T}_x} \mathrm{TT}_{x,t}. \tag{7}$$

Example 2 Let u_1, u_2, and u_3 be three receivers and let $C = \{1, 2, 5.5, 11\}$, $c_1 = 1$, $c_2 = 2$, and $c_3 = 11$. In this condition, u_1 can only use transmission rate of 1 and u_3 can use transmission rates of 1, 2, 5.5, and 11. When the sender multicasts a file x to the receivers, it considers four kinds of scheduling, as shown in Fig. 3. The scheduling t_1 is delivered to the receiver u_1 at a transmission rate of 1, and multicasts to the receivers u_2 and u_3 at a transmission rate of 2. When multicasting to the receivers u_2 and u_3, although the receiver u_3 can communicate at the transmission rate 11, the receiver u_3 is limited to a small transmission rate.

The sender selects one of 4 scheduling plans based on the multicast throughput $\mathrm{TH}_{x,t}$ or the multicast delay $\mathrm{TT}_{x,t}$. Table 2 shows $\mathrm{TH}_{x,t}$ and $\mathrm{TT}_{x,t}$ of each scheduling. From Table 2, t_4 is the best scheduling if the sender selects scheduling based on $\mathrm{TH}_{x,t}$. On the other hand, t_2 is the best scheduling if the sender selects scheduling based on $\mathrm{TT}_{x,t}$. ■

The throughput of Definition 1 cannot be evaluated considering the message size. Because the size of the delivered message is one of the important measures to evaluate the coded caching scheme, this paper focuses on the delay, which can be evaluated considering the message size.

In the following sections, we will refer to the scheduling with the smallest delay as the best scheduling. From (6), the best scheduling is the case of multicasting at once with a rate that is available to receivers. The rate is given by the minimum value among the maximum transmission rates of each receiver. In Example 2, the delay is the smallest when multicasting at the lowest transmission rate among c_1, c_2, and c_3.

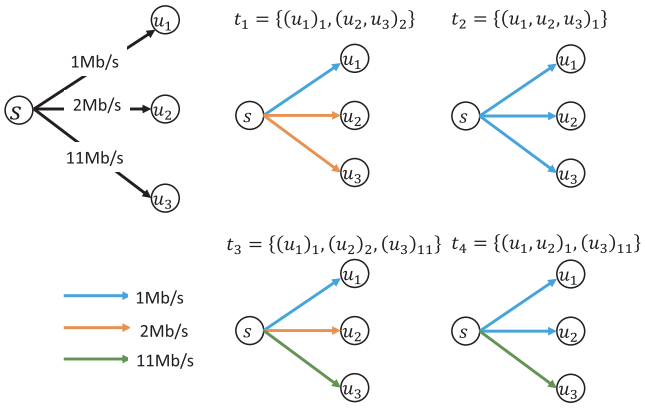


Fig. 3 Scheduling considered in Example 2.

Table 2 The multicast throughput $\mathrm{TH}_{x,t}$ and the multicast delay $\mathrm{TT}_{x,t}$ of each scheduling in Example 2.

Scheduling	Multicast throughput	Multicast delay
$t_1 = \{(u_1)_1, (u_2, u_3)_2\}$	2.5	1.5
$t_2 = \{(u_1, u_2, u_3)_1\}$	3	1 (best)
$t_3 = \{(u_1)_1, (u_2)_2, (u_3)_{11}\}$	4.67	1.59
$t_4 = \{(u_1, u_2)_1, (u_3)_{11}\}$	6.5 (best)	1.09

4. Coded Caching Scheme for Multi-Rate Wireless Networks

In this section, we define measures to evaluate the coded caching schemes and propose a new delivery scheme suitable for multi-rate wireless networks.

4.1 Evaluation Measures

The worst-case size of transmission messages normalized by file size is used as the measures of the coded caching schemes [2]. On the other hand, in the multi-rate wireless networks, the multicast delay is used as the measures of the scheduling. In this paper, we consider coded caching in the multi-rate environment. We newly define a coded caching delay for evaluating the coded caching schemes.

We consider the worst case where the size of delivery messages becomes the largest to estimate the delivery cost.

Definition 3 (The worst-case size of transmission messages). We define the worst-case size of the transmission messages as

$$R \triangleq \max_{d \in \mathcal{N}^K} R^d. \tag{8}$$

For the vector of request d, let $\mathcal{X} = \{x_u | u \in \mathcal{U}\}$ be the set of messages that are delivered by the coded caching scheme.

Table 3 The size of each message transmitted by the zero-padding scheme and its minimum delay in Example 3.

Coded message	Size	Delay	Best scheduling
x_{123}	0.222	0.222	$\{(u_1, u_2, u_3)_1\}$
x_{12}	0.111	0.056	$\{(u_1, u_2)_2\}$
x_{13}	0.222	0.222	$\{(u_1, u_3)_1\}$
x_{23}	0.222	0.222	$\{(u_2, u_3)_1\}$
x_1	0.111	0.010	$\{(u_1)_{11}\}$
x_2	0.111	0.056	$\{(u_2)_2\}$
x_3	0.111	0.111	$\{(u_3)_1\}$
Total	1.11	0.899	

The total size of the messages is given by

$$R^d = \sum_{x \in \mathcal{X}} |x|. \tag{9}$$

By definition, for any vector of requests d, the total size of the transmission messages is less than or equal to the size of the worst-case R so that the network load can be estimated.

In the coded caching scheme, the time until each receiver receives the data necessary for recovering the request is the time until all messages are transmitted since the coded messages are transmitted.

Definition 4 (Coded caching delay). Let $\mathcal{X} = \{x_u | u \in \mathcal{U}\}$ be the set of messages that are delivered by the coded caching scheme. The multi-rate multicast delay of each message $x \in \mathcal{X}$ is given by (7). The coded caching delay is defined as

$$\mathrm{TT} \triangleq \sum_{x \in \mathcal{X}} \mathrm{TT}_{x,\min}. \tag{10}$$

Example 3 Consider the same setting as Example 1, so $M_1 = 1$, $M_2 = 1.5$, and $M_3 = 2$. In addition, let $C = \{1, 2, 5.5, 11\}$ and suppose that the transmission rates available to each receiver are $c_1 = 11$, $c_2 = 2$, and $c_3 = 1$. The size of each message is shown in Table 3. Table 3 also shows the delay with the best scheduling selected based on the multi-rate multicast delay. For example, the size of x_{123} is 0.222, and the best scheduling to multicast x_{123} is $\{(u_1, u_2, u_3)_1\}$ from Example 2.

The bottom of Table 3 shows the total size of the messages and the total delay in the zero-padding scheme. In this example, the worst-case size of transmission messages is 1.11, and the coded caching delay is 0.899. ■

Consider a situation that a receiver requiring a large subfile can receive with a large transmission rate and another receiver requiring a small subfile can receive with a small transmission rate. Then, the zero-padding scheme transmits the coded message, which is an XOR of two subfiles after zero-padding to scale to large subfile, with the small transmission rate. The weakness of the zero-padding scheme is that the scheme sometimes sends redundant data.

Algorithm 2 Proposed delivery scheme

Delivery Phase
for $(k = K; k > 0; k--)$ **do**
  for each subset $\mathcal{U}$ of k receivers do **do**
    $\mathcal{U}_{\text{send}} \leftarrow \mathcal{U}$;
    while $\mathcal{U}_{\text{send}} \neq \phi$ **do**
      Minsize $\leftarrow \min_{i \in \mathcal{U}} |w_{d_i, \mathcal{U}\backslash\{i\}}|$;
      $x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}} \leftarrow$ all-zero vector of length Minsize bits;
      for $l \in \mathcal{U}_{\text{send}}$ **do**
        if $|w_{d_l, \mathcal{U}\backslash\{l\}}| =$ Minsize **then**
          add receiver l to $\mathcal{U}_{\text{temp}}$;
        end if
        $w_{d_l, \mathcal{U}\backslash\{l\}}^{\mathcal{U}_{\text{send}}} \leftarrow$ the first Minsize bits of $w_{d_l, \mathcal{U}\backslash\{l\}}$;
        $w_{d_l, \mathcal{U}\backslash\{l\}} \leftarrow$ the remaining bits of $w_{d_l, \mathcal{U}\backslash\{l\}}$;
        $x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}} \leftarrow x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}} \oplus w_{d_l, \mathcal{U}\backslash\{l\}}^{\mathcal{U}_{\text{send}}}$;
      end for
      Multicast the coded data $x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}}$ to receivers in $\mathcal{U}_{\text{send}}$.
      $\mathcal{U}_{\text{send}} \leftarrow \mathcal{U}_{\text{send}} \backslash \mathcal{U}_{\text{temp}}$;
    end while
  end for
end for

Recovery Phase of receiver k
for each subset $\mathcal{U}$ including k **do**
  $x_{\mathcal{U}} \leftarrow$ null vector;
  for each subfile $x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}}$ **do**
    $x_{\mathcal{U}} \leftarrow x_{\mathcal{U}} \| x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}}$;
  end for
  for $l \in \mathcal{U}\backslash\{k\}$ **do**
    $x_{\mathcal{U}} \leftarrow x_{\mathcal{U}} \overline{\oplus} w_{d_l, \mathcal{U}\backslash\{l\}}$;
  end for
  $w_{d_k, \mathcal{U}\backslash\{k\}} \leftarrow x_{\mathcal{U}}$;
end for
w_{d_k} is restored by all subfiles of w_{d_k};

Example 4 In Table 1, when x_{13} is transmitted, $w_{d_3,1}$ is padded with $|w_{d_1,3}| - |w_{d_3,1}|$ zeros because $|w_{d_1,3}|$ is larger than $|w_{d_3,1}|$. In this case, the receiver u_1 needs to use the small transmission rate $c_3 = 1$ to multicast in spite of the fact that u_1 can receive with the transmission rate $c_1 = 11 (> c_3)$. From a different viewpoint, the zero-padding scheme transmits inactive data, i.e. the part of zero-padding, to the receiver u_3 with the small transmission rate. ■

4.2 Proposed Delivery Scheme

For the problem described in the previous subsection, we consider XORing by dividing subfile according to a small subfile rather than by zero-padding to scale to a large subfile.

We propose Algorithm 2 as a method to make the coded message by XORing in accordance with a small subfile. The placement phase operates on the same algorithm as the zero-padding scheme. The delivery phase operates as follows. At first, we choose k receivers from K receivers to form a subset $\mathcal{U}$. $\mathcal{U}_{\text{send}}$ refers to the subset of receivers who receive the coded data. Then, we equalize the file size for encoding.

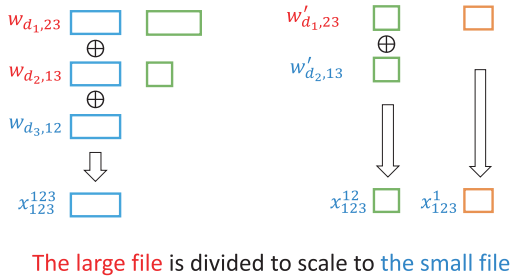


Fig. 4 An example of encoding using the proposed scheme.

Therefore, we find the smallest file size among the files that will be encoded, and then let Minsize be the smallest file size. After that, we encode the first Minsize bits of each file, and then we send the coded data $x_{\mathcal{U}}^{\mathcal{U}_{\text{send}}}$ to receivers in $\mathcal{U}_{\text{send}}$. The superscript of the coded message refers to the subset of receivers who receive it. Now, we do not need to send any additional data to the receivers who requested the subfile of size Minsize, so we remove such receivers from $\mathcal{U}_{\text{send}}$.

Example 5 We consider transmitting the equivalent data to x_{123} of Table 1 using the proposed scheme. At first, $w_{d_1,23}$ and $w_{d_2,13}$ are divided into two subfiles that the size of one subfile is equal to $|w_{d_3,12}|$ as shown in Fig. 4. Then, the remaining of $w_{d_1,23}$ is divided into two subfiles that the size of one subfile is equal to the remaining of $w_{d_2,13}$. After that, coded messages are made by XOR-ing subfiles of the same size. Then, x_{123}^{13} is transmitted to the receivers u_1, u_2, and u_3, x_{123}^{12} is transmitted to the receivers u_1 and u_2, and x_{123}^{1} is transmitted to the receiver u_1. ■

Example 6 After receiving the data in Table 4, receiver u_1 recovers tha requested file w_{d_1} as follows. u_1 receives x_{123}^{123}, x_{123}^{12}, x_{123}^{1}, x_{12}^{12}, x_{12}^{1}, x_{13}^{13}, x_{13}^{1}, and x_1. For $\mathcal{U} = \{1,2,3\}$, x_{123} is recovered by $x_{123} = x_{123}^{123} \| x_{123}^{12} \| x_{123}^{1}$ and $w_{d_1,23}$ is recovered by $w_{d_1,23} = x_{123} \oplus w_{d_2,13} \oplus w_{d_3,12}$, where $w_{d_2,13}$ and $w_{d_3,12}$ are stored in memory of receiver u_1. Similarly, u_1 recovers $w_{d_1,2}$, $w_{d_1,3}$, and $w_{d_1,\phi}$. Now that u_1 has all subfiles of w_{d_1}, so u_1 can recover w_{d_1}. ■

Example 7 Assume the same network setting as Example 3, that is, 3 receivers u_1, u_2, and u_3 have the memories of size $M_1 = 1$, $M_2 = 1.5$, and $M_3 = 2$ and can receive message with the transmission rates $c_1 = 11$, $c_2 = 2$, and $c_3 = 1$, respectively.

Table 4 shows the coded messages, and Table 5 shows the size of each message and the minimum delay. For example, when the equivalent data to x_{123} of Table 1 is transmitted, the proposed scheme transmits the messages x_{123}^{123}, x_{123}^{12}, and x_{123}^{1} with the scheduling $\{(u_1,u_2,u_3)_1\}$, $\{(u_1,u_2)_2\}$, and $\{(u_1)_{11}\}$, respectively. Then, the delay is $0.056 + 0.028 + 0.010 = 0.094$ with the proposed scheme, and it is 0.128 faster than the zero-padding scheme. The total size of the messages and the coded caching delay

Table 4 The messages transmitted by the proposed scheme in Example 6.

Coded message	Size
$x_{123}^{123} = w_{d_1,23}^{123} \oplus w_{d_2,13}^{123} \oplus w_{d_3,12}$	$\frac{M_1}{N}\frac{M_2}{N}\left(1-\frac{M_3}{N}\right)$
$x_{123}^{12} = w_{d_1,23}^{12} \oplus w_{d_2,13}^{12}$	$\frac{M_1}{N}\left(1-\frac{M_2}{N}\right)\frac{M_3}{N} - \lvert x_{123}^{123}\rvert$
$x_{123}^{1} = w_{d_1,23}^{1}$	$\left(1-\frac{M_1}{N}\right)\frac{M_2 M_3}{N N} - \lvert x_{123}^{123}\rvert - \lvert x_{123}^{12}\rvert$
$x_{12}^{12} = w_{d_1,2}^{12} \oplus w_{d_2,1}$	$\frac{M_1}{N}\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_{12}^{1} = w_{d_1,2}^{1}$	$\left(1-\frac{M_1}{N}\right)\frac{M_2}{N}\left(1-\frac{M_3}{N}\right) - \lvert x_{12}^{12}\rvert$
$x_{13}^{13} = w_{d_1,3}^{13} \oplus w_{d_3,1}$	$\frac{M_1}{N}\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_{13}^{1} = w_{d_1,3}^{1}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\frac{M_3}{N} - \lvert x_{13}^{13}\rvert$
$x_{23}^{23} = w_{d_2,3} \oplus w_{d_3,2}$	$\left(1-\frac{M_1}{N}\right)\frac{M_2}{N}\left(1-\frac{M_3}{N}\right)$
$x_{23}^{2} = w_{d_2,3}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\frac{M_3}{N} - \lvert x_{23}^{23}\rvert$
$x_1 = w_{d_1,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_2 = w_{d_2,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$
$x_3 = w_{d_3,\phi}$	$\left(1-\frac{M_1}{N}\right)\left(1-\frac{M_2}{N}\right)\left(1-\frac{M_3}{N}\right)$

Table 5 The size of each message transmitted by the proposed scheme and its minimum delay.

Coded message	Size	Delay	Best scheduling
x_{123}^{123}	0.056	0.056	$\{(u_1,u_2,u_3)_1\}$
x_{123}^{12}	0.056	0.028	$\{(u_1,u_2)_2\}$
x_{123}^{1}	0.111	0.010	$\{(u_1)_{11}\}$
x_{12}^{12}	0.056	0.028	$\{(u_1,u_2)_2\}$
x_{12}^{1}	0.056	0.005	$\{(u_1)_{11}\}$
x_{13}^{13}	0.056	0.056	$\{(u_1,u_3)_1\}$
x_{13}^{1}	0.167	0.015	$\{(u_1)_{11}\}$
x_{23}^{23}	0.111	0.111	$\{(u_2,u_3)_1\}$
x_{23}^{2}	0.111	0.056	$\{(u_2)_2\}$
x_1	0.111	0.010	$\{(u_1)_{11}\}$
x_2	0.111	0.056	$\{(u_2)_2\}$
x_3	0.111	0.111	$\{(u_3)_1\}$
Total	1.11	0.540	

are shown at the bottom of Table 5. In this example, the total size is 1.11, and the delay is 0.540. From the comparison with the zero-padding scheme, the total size of messages is equivalent, and the coded caching delay is 60% of the delay of the zero-padding scheme. ■

The proposed scheme improves the coded caching delay by avoiding sending useless data to the receivers with the small transmission rate. Moreover, the total size of messages of the proposed scheme is equal to that of the zero-padding scheme.

5. Numerical Results

We evaluate the performance of the coded caching scheme in the multi-rate wireless network by the data size and the coded caching delay. For the comparison, the un-

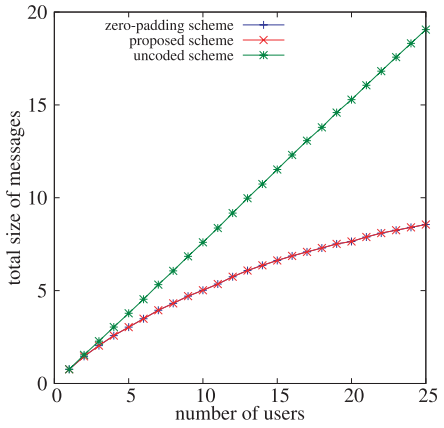


Fig. 5 Comparison of total size of messages.

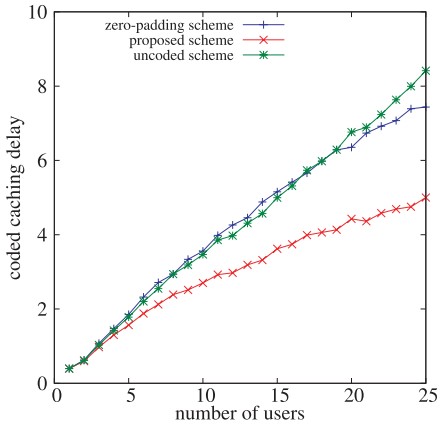


Fig. 6 Comparison of coded caching delay.

coded caching scheme described in [2] and the zero-padding scheme proposed in [7] are targets for the evaluation.

In the network model, we assume different memory size at each receiver. However, it is considered that the type of memory size is fixed in the actual network environment. In this experiment, we consider a scenario that the sender has $N = 100$ files of size 20 MB[†] and the receiver has a memory of size 100 MB, 300 MB, 500 MB, or 1000 MB[††]. Thus, we randomly choose the memory size of the receiver from $\{5, 15, 25, 50\}$. Also, we randomly choose the transmission rate from $\{1, 2, 5.5, 11\}$ independent of the memory size.

Figure 5 shows the comparison of the total size of transmitted messages. The size should be small to reduce the network load. Figure 6 shows the comparison of the coded caching delay. The delay should be small to satisfy the requests of the receivers quickly. Each result is the average of 20 trials with randomly changed memory size and transmission rate.

From Fig. 5, the size of the message transmitted by the proposed scheme is equal to that by the zero-padding scheme, and it is also smaller than that by the uncoded caching scheme. Therefore, the coded caching schemes can reduce the size of the message so that the network load will be smaller. As can be seen from Figs. 5 and 6, although the zero-padding scheme reduces the total size of messages from the uncoded scheme, the delay is almost the same to the delay of the uncoded scheme.

In the zero-padding scheme, the sender can transmit coded messages, which are smaller than the messages transmitted by the uncoded scheme, to target receivers at once by multicast. However the sender has to use the smallest transmission rate among the target receivers. Also, this scheme generates an inactive data by zero-padding to scale to the small subfile to the large subfile and transmits it with the small rate. On the other hand, in the uncoded scheme, the sender can transmit the messages to each receiver with as large as possible transmission rate. Due to the difference in available rates, the zero-padding scheme has a large delay regardless of the difference in the total size of messages. The proposed scheme improves the delay from both of the conventional schemes. If $K = 15$, the delay is shortened by about 30%. In the proposed scheme, the sender can transmit the coded message with as large as possible transmission rate because this scheme does not generate any inactive data by splitting the subfiles, unlike the zero-padding scheme.

For further comparison, we experimented with the following two patterns. One is a pattern in which users with a large memory size can communicate at a large transmission rate, i.e., $c_1 \leq c_2 \leq \cdots \leq c_k$. We call it pattern A. The other is a pattern in which users with a large memory size can communicate at a small transmission rate, i.e., $c_1 \geq c_2 \geq \cdots \geq c_k$. We call it pattern B. The results of each pattern are shown in Figs. 7 and 8. In pattern A, the proposed scheme and zero-padding scheme give the same results. It can be seen that the conditions unfavorable to the proposed scheme are not inferior to the zero padding scheme. In pattern B, the proposed scheme reduces the delay significantly from the existing schemes, and the zero-padding scheme is worse than the uncoded scheme. From the above results, the proposed scheme has a smaller delay than the zero-padding scheme for any condition.

In order to measure the practical delay time, it is necessary to consider the time taken for coding and decoding. For a large number of receivers, a large coding times is a problem in the coded caching schemes and beyond the scope of discussion in this paper. This problem has been discussed in [14]–[16]. In future works, these methods will be applied to proposed scheme and the conventional schemes and we will evaluate the partical delay time of these scheme in multi-rate wireless networks.

6. Conclusion

We discussed the coded caching problem in the multi-rate wireless network. We defined the coded caching delay to evaluate the coded caching scheme in this network. We discussed the weak point of the zero-padding scheme on the

[†]It is about 10 minutes by the animation for mobile devices.
[††]Most web browsers can resize the capacity of cache up to 1024 MB.

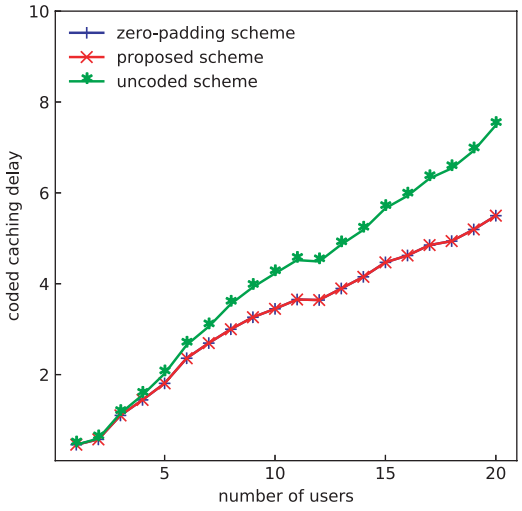


Fig. 7 Comparison of coded caching delay with pattern A.

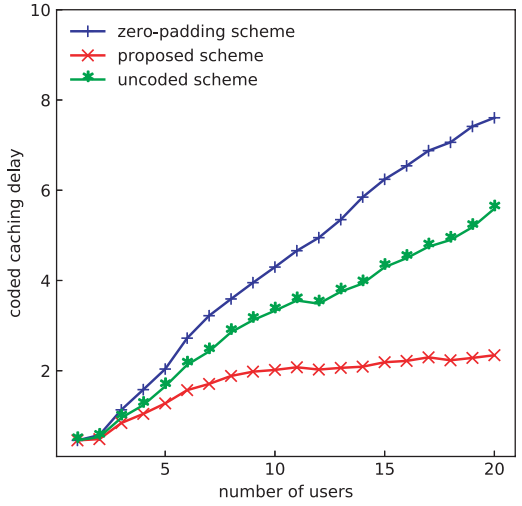


Fig. 8 Comparison of coded caching delay with pattern B.

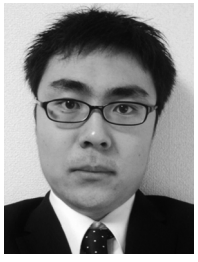
coded caching delay. We proposed the new coded caching scheme to solve the weak point and have evaluated it based on the coded caching delay. From the comparisons with the conventional schemes, we have shown that the delay is shortened by between 22% and 40% if the number of receivers is ten or more. Notably, the proposed scheme is suitable for multi-rate wireless networks.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP19K21534.

References

[1] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," Proc. IEEE INFOCOM, pp.1–9, San Diego, CA, March 2010.
[2] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," IEEE Trans. Inf. Theory, vol.60, no.5, pp.2856–2867, May 2014.
[3] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," IEEE/ACM Trans. Netw, vol.23, no.4, pp.1029–1040, April 2015.
[4] U. Niesen and M. Maddah-Ali, "Coded caching with nonuniform demands," Proc. 2014 IEEE INFOCOM WKSHPS, pp.221–226, Toronto, ON, April 2014.
[5] J. Hachem, N. Karamchandani, and S. Diggavi, "Effect of number of users in multi-level coded caching," Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT), pp.1701–1705, Hong Kong, June 2015.
[6] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," Proc. IEEE ITA, pp.98–107, San Diego, CA, Feb. 2015.
[7] S. Wang, W. Li, X. Tian, and H. Liu, "Coded Caching with Heterogenous Cache Sizes," arXiv:1504.01123v3, cs.IT, Aug. 2015.
[8] A.M. Ibrahim, A.A. Zewail, and A. Yener, "Centralized coded caching with heterogeneous cache sizes," Proc. 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp.1–6, San Francisco, CA, March 2017.
[9] J. Zhang, X. Lin, C.-C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT), pp.1686–1690, Hong Kong, June 2015.
[10] A. Daniel and W. Yu, "Optimization of heterogeneous coded caching," arXiv:1708.04322v1, cs.IT, Aug. 2017.
[11] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, and D. Gunduz, "Coded caching with heterogeneous cache sizes and link qualities: The two-user case," Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT), pp.1545–1549, Vail, CO, June 2018,
[12] C. Chou, A. Misra, and J. Qadir, "Low latency broadcast in multi-rate wireless mesh networks," University of New South Wales, School of Computer Science and Engineering Sydney, 2005.
[13] A. Ben Hassouna, H. Koubaa, and L.A. Saidane, "Multicast throughput subject to delay in multi-rate wireless networks," Proc. 2017 14th IEEE Annual Consumer Communications and Networking Conference (CCNC), pp.796–801, Las Vegas, NV, Jan. 2017.
[14] L. Tang and A. Ramamoorthy, "Low subpacketization schemes for coded caching," Proc. 2017 IEEE Int. Symp. Inf. Theory (ISIT), pp.2790–2794, Aachen, 2017.
[15] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," IEEE Trans. Inf. Theory, vol.63, no.9, pp.5821–5833, Sept. 2017.
[16] H.H. Suthan Chittoor, Bhavana M., and P. Krishnan, "Coded Caching via Projective Geometry: A new low subpacketization scheme," Proc. 2019 IEEE Int. Symp. Inf. Theory (ISIT), Paris, France, pp.682–686, 2019.
[17] M. Takita, M. Hirotomo, and M. Morii, "Coded caching in multi-rate wireless network," 2018 IEEE International Conference on Communications Workshops (ICC Workshops), pp.1–6, Kansas City, MO, May 2018.



Makoto Takita received his B.E., M.E., and D.E. degrees from Kobe University, Japan, in 2014, 2015, and 2018, respectively. In 2018, he was a Researcher at the Graduate School of Engineering, Kobe University, Japan. Since 2019, he has been an Assistant Professor at the School of Social Information Science, University of Hyogo, Japan. His research interests include coding theory, information networks, and information security.

Masanori Hirotomo received his B.E., M.E., and D.E. degrees from the University of Tokushima, Japan, in 2000, 2002, and 2006, respectively. From 2005 to 2006, he was a Research Associate in the Department of Intelligent Systems and Information Science, Faculty of Engineering at the University of Tokushima, Japan. From 2006 to 2008, he was a Researcher at the Hyogo Institute of Information Education Foundation, Japan. From 2008 to 2011, he was an Assistant Professor at the Graduate School of Engineering, Kobe University, Japan. From 2011 to 2013, he was an Assistant Professor at the Computer and Network Center, Saga University, Japan. Since 2013, he has been an Associate Professor at the Graduate School of Science and Engineering, Saga University, Japan. His research interests include coding theory and information security. He is a member of the IEEE.

Masakatu Morii received his B.E. degree in electrical engineering and his M.E. degree in electronics engineering from Saga University, Saga, Japan, and his D.E. degree in communication engineering from Osaka University, Osaka, Japan in 1983, 1985, and 1989, respectively. From 1989 to 1990, he was an Instructor in the Department of Electronics and Information Science, Kyoto Institute of Technology, Japan. From 1990 to 1995, he was an Associate Professor in the Department of Computer Science, Faculty of Engineering at Ehime University, Japan. From 1995 to 2005, he was a Professor in the Department of Intelligent Systems and Information Science, Faculty of Engineering, at the University of Tokushima, Japan. Since 2005, he has been a Professor in the Department of Electrical and Electronics Engineering, Faculty of Engineering, at Kobe University, Japan. His research interests include error correcting codes, cryptography, discrete mathematics, computer networks, and information security. He is a member of the IEEE.