POSITION PAPER

Overloaded MIMO Spatial Multiplexing Independent of Antenna Setups

Satoshi DENNO[†a)], ***Senior Member***, **Takumi SUGIMOTO**[††], **Koki MATOBA**[†††], ***Nonmembers***, ***and*** **Yafei HOU**[†], ***Senior Member***

SUMMARY This paper proposes overloaded MIMO spatial multiplexing that can increase the number of spatially multiplexed signal streams despite of the number of antennas on a terminal and that on a receiver. We propose extension of the channel matrix for the spatial multiplexing to achieve the superb multiplexing performance. Precoding based on the extended channel matrix plays a crucial role in carrying out such spatial multiplexing. We consider three types of QR-decomposition techniques for the proposed spatial multiplexing to improve the transmission performance. The transmission performance of the proposed spatial multiplexing is evaluated by computer simulation. The simulation reveals that the proposed overloaded MIMO spatial multiplexing can implement 6 stream-spatial multiplexing in a 2×2 MIMO system, i.e., the overloading ratio of 3.0. The superior transmission performance is achieved by the proposed overloaded MIMO spatial multiplexing with one of the QR-decomposition techniques.

key words: *overloaded MIMO, spatial multiplexing, QR-decomposition, precoding, overloading ratio*

1. Introduction

Communication speed has been raised to about Gbps by using many cutting-edge techniques even in wireless communication systems. Among them, multiple input multiple output (MIMO) spatial multiplexing has been playing an important role in enhancing communication speed [1]–[3]. For the enhancement, many MIMO techniques have been proposed such as serial interference cancelers based on the minimum mean square error (MMSE), precoders, iterative decoders, and so on [4]–[7]. To multiply the throughput enhancement, lots of antennas are installed on the base station in the fifth generation (5G) cellular system, which is called "Massive MIMO" [8]–[10]. While those techniques achieve superior performance [11]–[13], those techniques are unable to increase the user throughput. For increasing the user throughput, some techniques have been proposed such as non-orthogonal multiple access [14]–[19], faster-than-Nyquist (FTN) [20], and overloaded MIMO spatial multiplexing [21]. Especially, overloaded MIMO spatial multiplexing can increase the download throughput in massive MIMO systems by making use of the system configuration where lots of antennas are installed on the base stations. The use of massive MIMO in the latest wireless systems such as the 5G cellular system has induced researchers to have an interest in overloaded MIMO spatial multiplexing. Since the number of the spatially multiplexed signal streams is set to be more than the degree of freedom of linear receivers in overloaded MIMO systems, non-linear receivers have been considered for the signal detection at the receivers [22], [23]. However, although non-linear receivers achieve superior transmission performance, because the non-linear receivers execute the brute force search, they imposes prohibitive high computational complexity on the receivers. Complexity reduction techniques for them have been proposed [24]–[27]. Some techniques to improve the performance of those complexity reduced non-linear detectors have also been investigated [28], [29]. On the other hand, linear detectors have been considered for overloaded MIMO spatial multiplexing where virtual channels are introduced to assist signal detection in overloaded MIMO channels [30], [31]. Those techniques need complex signal processing, such as oversampling and non-linear optimization, in addition to the linear signal detection and the linear precoding. For further complexity reduction, linear detectors based on serial interference cancellation have been proposed to detect overloaded MIMO spatial multiplexed signal streams [32], [33]. Whereas many different techniques have been utilized for overloaded MIMO systems, they are proposed for the systems where the number of receive antennas is less than that of transmit antennas. Especially, the number of spatially multiplexed signal streams is limited by the number of transmit antennas in all of the conventional overloaded MIMO systems, including the systems with the virtual channels.

This paper proposes overloaded MIMO spatial multiplexing that can increase the number of spatially multiplexed signal streams to more than that of not only transmit antennas but also receive antennas in order to enhance link throughput. Actually, the number of spatially multiplexed signal streams can be raised despite of that of antenna settings on a receiver or a transmitter*. For such spatial multiplexing, the channel matrix is extended with some appropriate matri-

Manuscript received December 21, 2023.
Manuscript revised May 11, 2024.
Manuscript publicized August 1, 2024.

[†]Faculty of Environmental, Life, Natural Science and Technology, Okayama University, Okayama-shi, 700-8530 Japan.
[††]Graduate School of Natural Science and Technology, Okayama University, Okayama-shi, 700-8530 Japan.
[†††]Graduate School of Environmental, Life, Natural Science and Technology, Okayama University, Okayama-shi, 700-8530 Japan.
a) E-mail: denno@okayama-u.ac.jp

DOI: 10.23919/transcom.2023EBN0001

*While the proposed overloaded MIMO spatial multiplexing introduces a virtual transmission signal vector as shown in the following, the proposed multiplexing never makes use of the virtual channels.

ces in the proposed overloaded MIMO spatial multiplexing. Precoding based on the extended channel matrix is applied in order to make the number of the spatially multiplexed signal streams exceed that of the antennas on the terminal or the receiver, which characterizes the proposed overloaded MIMO spatial multiplexing, while the number of the signal streams is limited by the number of antennas in conventional overloaded MIMO systems. Three types of QR-decomposition techniques are considered for the precoding to improve the transmission performance. The performance of the proposed overloaded MIMO spatial multiplexing is evaluated by computer simulation.

Throughout the paper, $\Re[c]$ and $\Im[c]$ represent a real part and an imaginary part of a complex number c, respectively. Superscript T and H indicate transpose and Hermitian transpose of a matrix or a vector, respectively. tr[$\mathbf{A}$], $\mathbf{A}_{m,n}$, and $\mathrm{E}[c]$ indicate a trace of a matrix $\mathbf{A}$, an (m,n)-entry of a matrix $\mathbf{A}$, and the ensemble average of c.

2. System Model

We assume an MIMO system where a transmitter with N_T antennas transmits some signal streams for a receiver with N_R antennas. While channel coding is usually used in current wireless communication systems, any channel coding is not assumed in the system. The signal streams from the modulators are provided to a precoder that is explained in the following section. The precoder output signal streams are fed to the N_T antennas for the signal transmission. The signal streams are traveling through fading channels and received at the N_R antennas on the receiver. Let $\mathbf{Y} \in \mathbb{C}^{N_\mathrm{R}}$ represent a received signal vector, the system model is written as follows.

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{X} \in \mathbb{C}^{N_\mathrm{T}}$, $\mathbf{N} \in \mathbb{C}^{N_\mathrm{R}}$, and $\mathbf{H} \in \mathbb{C}^{N_\mathrm{R}\times N_\mathrm{T}}$ denote a transmission signal vector, an additive white Gaussian noise (AWGN) vector, and a channel matrix defined as,

$$\mathbf{H} = \begin{pmatrix} h(1,1) & h(1,2) & \cdots & h(1,N_\mathrm{T}) \\ h(2,1) & h(2,2) & & \vdots \\ \vdots & & \ddots & \\ h(N_\mathrm{R},1) & \cdots & & h(N_\mathrm{R},N_\mathrm{T}) \end{pmatrix}. \tag{2}$$

In (2), $h(n,m) \in \mathbb{C}$ represents a channel impulse response between the mth antenna on the transmitter and the nth antenna on the receiver, respectively. The system model is illustrated in the Fig. 1. This system is regarded as one of single-user MIMO systems.

In conventional MIMO systems, the number of spatially multiplexed signal streams N_S is reduced to $\min[N_\mathrm{T}, N_\mathrm{R}]$. As is known, overload MIMO techniques increase the number of signal streams to N_T even when the number of receive antennas N_R is less than that of transmit antennas N_T. The number of the signal streams is restricted by the number of the antennas on a receiver or a transmitter.

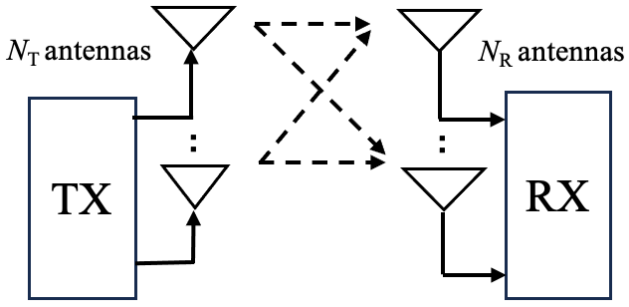


Fig. 1 System model.

We propose a technique to overcome the restriction in this paper. In a word, the proposed technique enables the number of spatially multiplexed signal streams N_S to be greater than the maximum of the number of the antennas on receiver and that on transmitter, i.e., $N_\mathrm{S} > \max[N_\mathrm{T}, N_\mathrm{R}]$.

3. Overloaded MIMO Spatial Multiplexing

While the system model is written in (1), the system model can be rewritten as follows.

$$\overline{\mathbf{Y}} = \begin{pmatrix} \mathbf{H} & \mathbf{0}_{N_\mathrm{R}\times N_\mathrm{B}} \\ \mathbf{0}_{N_\mathrm{A}\times N_\mathrm{T}} & \mathbf{0}_{N_\mathrm{A}\times N_\mathrm{B}} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \tilde{\mathbf{X}} \end{pmatrix} + \begin{pmatrix} \mathbf{N} \\ \mathbf{0}_{N_\mathrm{A}} \end{pmatrix} \tag{3}$$

In (3), $\mathbf{0}_{N\times M}$, $\tilde{\mathbf{X}} \in \mathbb{C}^{N_\mathrm{B}\times 1}$, and $\overline{\mathbf{Y}} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ denote an $N \times M$-dimensional null matrix, an N_B-dimensional virtual transmission signal vector, and an extended received signal vector defined as $\overline{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y}^\mathrm{T} & \mathbf{0}_{N_\mathrm{A}}^\mathrm{T} \end{pmatrix}^\mathrm{T}$ where $\mathbf{0}_N$ represents the N-dimensional null vector. We introduce an extended channel matrix $\overline{\mathbf{H}} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times(N_\mathrm{T}+N_\mathrm{B})}$ in this paper, which is defined as,

$$\overline{\mathbf{H}} = \begin{pmatrix} \mathbf{H} & \mathbf{J}_1 \\ \mathbf{J}_2 & \mathbf{J}_3 \end{pmatrix}. \tag{4}$$

In (4), $\mathbf{J}_1 \in \mathbb{C}^{N_\mathrm{R}\times N_\mathrm{B}}$, $\mathbf{J}_2 \in \mathbb{C}^{N_\mathrm{A}\times N_\mathrm{T}}$, and $\mathbf{J}_3 \in \mathbb{C}^{N_\mathrm{A}\times N_\mathrm{B}}$ indicate matrices, which are shown as examples in the following section. Let $\overline{\mathbf{X}} \in \mathbb{C}^{(N_\mathrm{T}+N_\mathrm{B})\times 1}$ denote an extended transmission signal vector defined as $\overline{\mathbf{X}} = \begin{pmatrix} \mathbf{X}^\mathrm{T} & \tilde{\mathbf{X}}^\mathrm{T} \end{pmatrix}^\mathrm{T}$, the system model can be further rewritten in the following.

$$\overline{\mathbf{Y}} = \overline{\mathbf{H}}\,\overline{\mathbf{X}} + \overline{\mathbf{N}}, \tag{5}$$

where $\overline{\mathbf{N}} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ represents an extended noise vector defined as follows.

$$\overline{\mathbf{N}} = \begin{pmatrix} \left(\mathbf{N} - \mathbf{J}_1\tilde{\mathbf{X}}\right)^\mathrm{T} & \left(-\mathbf{J}_2\mathbf{X} - \mathbf{J}_3\tilde{\mathbf{X}}\right)^\mathrm{T} \end{pmatrix}^\mathrm{T} \tag{6}$$

While the system defined in (1) comprises the transmitter with N_T antennas and the receiver with N_R antennas, the system in (5) looks extended to consist of the transmitter with $N_\mathrm{T} + N_\mathrm{B}$ antennas and the receiver with $N_\mathrm{R} + N_\mathrm{A}$ antennas. In other words, not only the number of the transmit antennas but also that of the receive antennas are increased in the extended system model. This extended system model enables overloaded MIMO spatial multiplexing independent

of antennas settings, in principle. Moreover, we propose a linear overloaded MIMO system, which can be implemented with small computational complexity. The detail is described in the following sections.

3.1 QR-Decomposition For Extended Channel Matrix

Triangulation of a channel matrix is introduced in not only wireless communication systems but also wired communication systems in order to improve the transmission performance or to reduce computational complexity. The triangulation can be performed not only at a receiver but also at a transmitter, because linear signal processing such as the triangulation can be transferred between a receiver and a transmitter in linear communication systems. The triangulation is usually implemented with QR-decomposition based on some signal processing techniques such as Gram-Schmidt orthonormalization. When the QR-decomposition is applied to the extended channel matrix, the channel matrix is decomposed into a semi-orthogonal matrix $\mathbf{Q}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$, a diagonal matrix $\mathbf{\Gamma}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$, and an upper triangular matrix $\mathbf{R}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$, which is expressed in the following.

$$\overline{\mathbf{H}}\mathbf{S}_{\mathrm{R}} = \mathbf{Q}_{\mathrm{R}}\mathbf{\Gamma}_{\mathrm{R}}\mathbf{R}_{\mathrm{R}} \tag{7}$$

However, the diagonal elements in the upper triangular matrix $\mathbf{R}_{\mathrm{R}}$ are all ones. $\mathbf{S}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$ in (7) denotes a transform matrix. For successful triangulation, the extended channel matrix has to be slim, which results in the following requirement.

$$N_{\mathrm{R}} + N_{\mathrm{A}} \geq N_{\mathrm{T}} + N_{\mathrm{B}} \tag{8}$$

When the requirement in (8) is not satisfied, we consider to apply triangulation to the Hermite transpose of the extended channel matrix $\overline{\mathbf{H}}$ as follows.

$$\overline{\mathbf{H}}^{\mathrm{H}}\mathbf{S}_{\mathrm{T}} = \mathbf{Q}_{\mathrm{T}}\mathbf{\Gamma}_{\mathrm{T}}\mathbf{R}_{\mathrm{T}} \tag{9}$$

Similar as the triangulation in (7), $\mathbf{S}_{\mathrm{T}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times(N_{\mathrm{R}}+N_{\mathrm{A}})}$, $\mathbf{Q}_{\mathrm{T}} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times(N_{\mathrm{R}}+N_{\mathrm{A}})}$, $\mathbf{\Gamma}_{\mathrm{T}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times(N_{\mathrm{R}}+N_{\mathrm{A}})}$, and $\mathbf{R}_{\mathrm{T}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times(N_{\mathrm{R}}+N_{\mathrm{A}})}$ denote a transform matrix, a semi-orthogonal matrix, a diagonal matrix, and an upper triangular matrix with ones in the diagonal positions. Same to (7), the Hermite transform of the extended channel matrix has to be slim, which can be expressed in the following equation.

$$N_{\mathrm{R}} + N_{\mathrm{A}} \leq N_{\mathrm{T}} + N_{\mathrm{B}} \tag{10}$$

As is shown above, the transform matrices are required for the QR-decomposition, which is truly necessary for the linear precoding and the THP as described in the following sections.

3.2 QR-Decomposition Techniques

Whereas the Gram-Schmidt orthonormalization is one of QR-decomposition techniques, other QR-decomposition techniques have been applied to MIMO systems. Some representatives of them are listed below. The notation of $\mathbf{S}_{\Omega}$ Ω = R or T is used for describing $\mathbf{S}_{\mathrm{R}}$ or $\mathbf{S}_{\mathrm{T}}$. The same notation is applied for $\mathbf{\Gamma}_{\mathrm{R}}$ or $\mathbf{\Gamma}_{\mathrm{T}}$ and $\mathbf{R}_{\mathrm{R}}$ or $\mathbf{R}_{\mathrm{T}}$ in the following.

(a) Sorted QR-decomposition [34]
This technique tries to arrange the diagonal elements of the diagonal matrix $\mathbf{\Gamma}_{\Omega}$ in ascending order, although it is not guaranteed to carry out the arrangement. In other words, let $\mathbf{\Gamma}_{\Omega}$ be defined as $\mathbf{\Gamma}_{\Omega} = \mathrm{diag}\left[\gamma_{\Omega}(1)\cdots\gamma_{\Omega}(N_{\Omega})\right]$ where $\gamma_{\Omega}(m) \in \mathbb{R}$ denotes the mth diagonal elements of the matrix $\mathbf{\Gamma}_{\Omega}$, the arrangement is mathematically described as,

$$|\gamma_{\Omega}(1)| \lesssim |\gamma_{\Omega}(2)| \lesssim \cdots \lesssim |\gamma_{\Omega}(N_{\Omega})|. \tag{11}$$

In the sorted QR-decomposition (SQRD), $\mathbf{S}_{\Omega}$ is reduced to a permutation matrix.

(b) Lattice Reduction [35]
While several techniques to implement the lattice reduction have been proposed and evaluated, the Lenstra–Lenstra–Lovász (LLL) algorithm [36] has been widely used in the field of communications[†], because it takes only a polynomial time length to find a near optimum vector from the view point of the lattice reduction[††]. Let the upper triangular matrix $\mathbf{R}_{\Omega}$ be defined as,

$$\mathbf{R}_{\Omega} = \begin{pmatrix} 1 & r_{\Omega}(1,2) & \cdots & r_{\Omega}(1,N_{\Omega}) \\ 0 & 1 & r_{\Omega}(2,3) & \vdots \\ \vdots & 0 & \ddots & r_{\Omega}(N_{\Omega}-1,N_{\Omega}) \\ 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{12}$$

The LLL algorithm achieves the following equations.

$$\Re\left[r_{\Omega}(n,m)\right] \leq \frac{1}{2}, \quad \Im\left[r_{\Omega}(n,m)\right] \leq \frac{1}{2} \tag{13}$$

$$\delta\,|\gamma_{\Omega}(m-1)|^2 \leq |\gamma_{\Omega}(m)|^2 + |\gamma_{\Omega}(m-1)\,r_{\Omega}(m-1,m)|^2 \tag{14}$$

$\delta \in \mathbb{R}$ in (14) denotes a parameter, which is set as $\frac{1}{2} < \delta < 1$ [37].

(c) Equal Gain Transform [38]
The equal gain transform [38] equalizes the diagonal elements with the transform matrix as,

$$\gamma_{\Omega}(1) = \gamma_{\Omega}(2) = \cdots = \gamma_{\Omega}(N_{\Omega}), \tag{15}$$

The matrix $\mathbf{S}_{\Omega}$ is served as a unitary matrix, i.e., $\mathbf{S}_{\Omega}^{\mathrm{H}}\mathbf{S}_{\Omega} = \mathbf{I}_{N_{\Omega}}$.

The derivation of those algorithms is described in the papers [34], [35], [38].

We propose overloaded MIMO spatial multiplexing with QR-decomposition where the above three techniques

[†]The LLL algorithm has been applied to overloaded MIMO systems, and its superior performance has been shown [32], [33]. This is the reason why the LLL algorithm is applied to our proposed overloaded MIMO spatial multiplexing.

[††]The LLL algorithm can be applied to any matrix as far as the matrix can be successfully QR-decomposed, i.e., the matrix is slim.

are applied in the following section[†].

3.2.1 Linear Precoding

If the above QR-decomposition techniques are applied to the extended channel matrix defined in (4), we can realize that the transform matrix $\mathbf{S}_\mathrm{R}$ can be used as a linear precoding matrix.

$$\overline{\mathbf{X}} = g_\mathrm{R}\mathbf{S}_\mathrm{R}\mathbf{D}, \tag{16}$$

where $\mathbf{D} \in \mathbb{C}^{(N_\mathrm{T}+N_\mathrm{B})\times 1}$ and $g_\mathrm{R} \in \mathbb{R}$ represent a modulation signal vector and a normalization factor that keeps the transmission power constant. Let the transform matrix $\mathbf{S}_\mathrm{R}$ be decomposed as $\mathbf{S}_\mathrm{R} = \begin{pmatrix}\mathbf{S}_{\mathrm{R},1}^\mathrm{T} & \mathbf{S}_{\mathrm{R},2}^\mathrm{T}\end{pmatrix}^\mathrm{T}$ where $\mathbf{S}_{\mathrm{R},1} \in \mathbb{C}^{N_\mathrm{T}\times(N_\mathrm{T}+N_\mathrm{B})}$ and $\mathbf{S}_{\mathrm{R},2} \in \mathbb{C}^{N_\mathrm{B}\times(N_\mathrm{T}+N_\mathrm{B})}$ denote an upper and a lower part of the transform matrix $\mathbf{S}_\mathrm{R}$, the transmission signal vector $\mathbf{X}$ and the virtual transmission signal vector $\tilde{\mathbf{X}}$ can be written as,

$$\mathbf{X} = g_\mathrm{R}\mathbf{S}_{\mathrm{R},1}\mathbf{D} \quad \text{and} \quad \tilde{\mathbf{X}} = g_\mathrm{R}\mathbf{S}_{\mathrm{R},2}\mathbf{D}. \tag{17}$$

The normalization factor g_R can be defined with only the transmission signal vector $\mathbf{X}$ because the vector is actually transmitted from the antennas.

$$g_\mathrm{R} = \sqrt{\frac{P_0}{\sigma_\mathrm{d}^2 \mathrm{tr}\left[\mathbf{S}_{\mathrm{R},1}^\mathrm{H}\mathbf{S}_{\mathrm{R},1}\right]}} = \sqrt{\frac{P_0}{N_\mathrm{T}\sigma_\mathrm{d}^2}} \tag{18}$$

$P_0 \in \mathbb{R}$ and $\sigma_\mathrm{d}^2 \in \mathbb{R}$ in (18) represent power of the transmission signal and that of the modulation signal defined as $\mathrm{E}\left[\mathbf{D}\mathbf{D}^\mathrm{H}\right] = \sigma_\mathrm{d}^2\mathbf{I}_{N_\mathrm{T}+N_\mathrm{B}}$.

In order to apply the above linear precoding, the system has to satisfy the requirement written in (8). When the system does not meet the requirement, another precoding that meets the requirement written in (10) is necessary, which is proposed in the next section.

3.2.2 THP Based on MMSE

Since precoding based on the MMSE criterion achieves superior transmission performance [39], we introduce precoding based on the MMSE for the system defined in (5). Let $\mathbf{A} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ represent a precoder input signal vector, a precoder based on the MMSE provides an extended transmission signal vector $\overline{\mathbf{X}}$ defined as,

$$\begin{aligned}\overline{\mathbf{X}} &= g_\mathrm{T}\overline{\mathbf{H}}^\mathrm{H}\left(\overline{\mathbf{H}}\,\overline{\mathbf{H}}^\mathrm{H}\right)^{-1}\mathbf{A} \\ &= g_\mathrm{T}\overline{\mathbf{H}}^\mathrm{H}\mathbf{S}_\mathrm{T}\left(\{\overline{\mathbf{H}}^\mathrm{H}\mathbf{S}_\mathrm{T}\}^\mathrm{H}\{\overline{\mathbf{H}}^\mathrm{H}\mathbf{S}_\mathrm{T}\}\right)^{-1}\mathbf{S}_\mathrm{T}^\mathrm{H}\mathbf{A}.\end{aligned} \tag{19}$$

In (19), $g_\mathrm{T} \in \mathbb{R}$ represents a normalization factor. In the above equation, the extended channel matrix is transformed with the transform matrix $\mathbf{S}_\mathrm{T}$. Since we assume that the extended channel matrix satisfies the requirement in (10) as is inferred at the end of the previous section, the extended channel matrix can be transformed in a manner defined in (9). Besides, the precoder input signal $\mathbf{A}$ is defined with a vector $\dot{\mathbf{D}} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ as $\mathbf{A} = \mathbf{S}_\mathrm{T}\dot{\mathbf{D}}$ where the vector $\dot{\mathbf{D}}$ is defined below. For the definition of the transmission signal vector $\mathbf{X}$, the semi-orthogonal matrix $\mathbf{Q}_\mathrm{T}$ is decomposed as $\mathbf{Q}_\mathrm{T} = \begin{pmatrix}\mathbf{Q}_{\mathrm{T},1}^\mathrm{T} & \mathbf{Q}_{\mathrm{T},2}^\mathrm{T}\end{pmatrix}^\mathrm{T}$ where $\mathbf{Q}_{\mathrm{T},1} \in \mathbb{C}^{N_\mathrm{T}\times(N_\mathrm{R}+N_\mathrm{A})}$ and $\mathbf{Q}_{\mathrm{T},2} \in \mathbb{C}^{N_\mathrm{B}\times(N_\mathrm{R}+N_\mathrm{A})}$ denote an upper matrix and a lower part of the semi-orthogonal matrix. If the term in the left hand side of (9) is substituted for (19), the transmission signal vector $\mathbf{X}$ and the virtual transmission signal vector $\tilde{\mathbf{X}}$ can be obtained by introducing the semi-orthogonal matrix $\mathbf{Q}_\mathrm{T}$ into (19) as follows.

$$\mathbf{X} = g_\mathrm{T}\mathbf{Q}_{\mathrm{T},1}\mathbf{\Gamma}_\mathrm{T}^{-1}\mathbf{R}_\mathrm{T}^{-\mathrm{H}}\dot{\mathbf{D}} = g_\mathrm{T}\mathbf{Q}_{\mathrm{T},1}\mathbf{\Gamma}_\mathrm{T}^{-1}\mathbf{V} \tag{20}$$

$$\tilde{\mathbf{X}} = g_\mathrm{T}\mathbf{Q}_{\mathrm{T},2}\mathbf{\Gamma}_\mathrm{T}^{-1}\mathbf{R}_\mathrm{T}^{-\mathrm{H}}\dot{\mathbf{D}} = g_\mathrm{T}\mathbf{Q}_{\mathrm{T},2}\mathbf{\Gamma}_\mathrm{T}^{-1}\mathbf{V} \tag{21}$$

In (20) and (21), $\mathbf{V} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ denotes a feedback filter output vector defined in the following.

$$\mathbf{R}_\mathrm{T}^\mathrm{H}\mathbf{V} = \dot{\mathbf{D}} \tag{22}$$

$\dot{\mathbf{D}} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ in (22) represents a modulation signal vector added with a Gaussian integer multiples vector $\mathbf{K}M \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ where $\mathbf{K} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times 1}$ and $M \in \mathbb{Z}$ indicate a Gaussian integer vector and a modulus. In a word, the vector $\dot{\mathbf{D}}$ is defined as $\dot{\mathbf{D}} = \mathbf{D}+\mathbf{K}M$. The entries of the Gaussian integer vector are successfully obtained as is done in the Tomlinson Harashima precoding (THP) [40][††]. Therefore, the proposed feedback filter is called as THP based on the MMSE even in this paper. Same to the linear precoding, since only the vector $\mathbf{X}$ is actually transmitted, the normalization factor g_T is obtained as,

$$g_\mathrm{T} = \sqrt{\frac{P_0}{\mathrm{tr}\left[\mathbf{Q}_{\mathrm{T},1}^\mathrm{H}\mathbf{Q}_{\mathrm{T},1}\mathbf{\Gamma}_\mathrm{T}^{-1}\mathbf{\Phi}_\mathrm{V}\mathbf{\Gamma}_\mathrm{T}^{-1}\right]}}. \tag{23}$$

In (23), $\mathbf{\Phi}_\mathrm{V} \in \mathbb{C}^{(N_\mathrm{R}+N_\mathrm{A})\times(N_\mathrm{R}+N_\mathrm{A})}$ represents an autocorrelation matrix of the feedback filter output vector $\mathbf{V}$, which is defined as $\mathbf{\Phi}_\mathrm{V} = \mathrm{E}\left[\mathbf{V}\mathbf{V}^\mathrm{H}\right]$.

3.3 SNR Performance

Since the transmission performance can be characterized by the signal to noise power ratio (SNR) of detector output signals in any wireless systems, the SNR performance is analyzed for the performance evaluation of the proposed overloaded MIMO spatial multiplexing in the following.

3.3.1 Linear Precoding

When the linear precoding defined in (16) is applied in the

[†]While the QR-decomposition based on the SQRD has been applied in detectors [34], the QR-decomposition can be transferred from a receiver to a transmitter, because our proposed overloaded MIMO spatial multiplexing is regarded as a linear system as described above.

[††]The triangular matrix $\mathbf{R}_\mathrm{T}$ obtained by the transform matrix $\mathbf{Q}_\mathrm{T}$ makes the feedback filter available at the transmitter.

$$\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{R}}\,\underline{\mathbf{N}}_{\mathrm{R}}^{\mathrm{H}}\right] = g_{\mathrm{R}}^{-2}\mathbf{Q}_{\mathrm{R}}^{\mathrm{H}}\mathrm{E}\left[\overline{\mathbf{N}}\,\overline{\mathbf{N}}^{\mathrm{H}}\right]\mathbf{Q}_{\mathrm{R}} = \mathbf{Q}_{\mathrm{R}}^{\mathrm{H}}\begin{pmatrix} \dfrac{\sigma^2}{g_{\mathrm{R}}^2}\mathbf{I}_{N_{\mathrm{R}}} + \sigma_{\mathrm{d}}^2\mathbf{J}_1\mathbf{S}_{\mathrm{R},2}\mathbf{S}_{\mathrm{R},2}^{\mathrm{H}}\mathbf{J}_1^{\mathrm{H}} & \sigma_{\mathrm{d}}^2\mathbf{J}_1\mathbf{S}_{\mathrm{R},2}\left(\mathbf{J}_2\mathbf{S}_{\mathrm{R},1}+\mathbf{J}_3\mathbf{S}_{\mathrm{R},2}\right)^{\mathrm{H}} \\ \sigma_{\mathrm{d}}^2\left(\mathbf{J}_2\mathbf{S}_{\mathrm{R},1}+\mathbf{J}_3\mathbf{S}_{\mathrm{R},2}\right)\left(\mathbf{J}_1\mathbf{S}_{\mathrm{R},2}\right)^{\mathrm{H}} & \sigma_{\mathrm{d}}^2\left(\mathbf{J}_2\mathbf{S}_{\mathrm{R},1}+\mathbf{J}_3\mathbf{S}_{\mathrm{R},2}\right)\left(\mathbf{J}_2\mathbf{S}_{\mathrm{R},1}+\mathbf{J}_3\mathbf{S}_{\mathrm{R},2}\right)^{\mathrm{H}} \end{pmatrix}\mathbf{Q}_{\mathrm{R}} \tag{25}$$

$$\begin{aligned}\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{T}}\,\underline{\mathbf{N}}_{\mathrm{T}}^{\mathrm{H}}\right] &= g_{\mathrm{T}}^{-2}\mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\mathrm{E}\left[\overline{\mathbf{N}}\,\overline{\mathbf{N}}^{\mathrm{H}}\right]\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\begin{pmatrix} \dfrac{\sigma^2}{g_{\mathrm{T}}^2}\mathbf{I}_{N_{\mathrm{R}}} + \mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\Phi_{\mathrm{V}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\mathbf{Q}_{\mathrm{T},2}^{\mathrm{H}}\mathbf{J}_1^{\mathrm{H}} & \mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\Phi_{\mathrm{V}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} \\ \left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\Phi_{\mathrm{V}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\left(\mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} & \left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\Phi_{\mathrm{V}}\boldsymbol{\Gamma}_{\mathrm{T}}^{-1}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} \end{pmatrix}\mathbf{S}_{\mathrm{T}} \\ &= \mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\begin{pmatrix} \dfrac{\sigma^2}{g_{\mathrm{T}}^2}\mathbf{I}_{N_{\mathrm{R}}} + \dfrac{M^2}{6}\mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\mathbf{Q}_{\mathrm{T},2}^{\mathrm{H}}\mathbf{J}_1^{\mathrm{H}} & \dfrac{M^2}{6}\mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} \\ \dfrac{M^2}{6}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\left(\mathbf{J}_1\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} & \dfrac{M^2}{6}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\left(\mathbf{J}_2\mathbf{Q}_{\mathrm{T},1}+\mathbf{J}_3\mathbf{Q}_{\mathrm{T},2}\right)^{\mathrm{H}} \end{pmatrix}\mathbf{S}_{\mathrm{T}}\end{aligned} \tag{28}$$

extended system, a detector input vector $\mathbf{Z} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times 1}$ is obtained with the transform matrix and the normalization factor, which can be fed to detectors.

$$\mathbf{Z}_{\mathrm{R}} \;=\; g_{\mathrm{R}}^{-1}\mathbf{Q}_{\mathrm{R}}^{\mathrm{H}}\overline{\mathbf{Y}} = \boldsymbol{\Gamma}_{\mathrm{R}}\mathbf{R}_{\mathrm{R}}\mathbf{D} + g_{\mathrm{R}}^{-1}\mathbf{Q}_{\mathrm{R}}^{\mathrm{H}}\overline{\mathbf{N}} \tag{24}$$

Because the matrix $\mathbf{R}_{\mathrm{R}}$ is upper triangular as shown in the above equation, serial interference cancelers (SICs) can be used to detect the modulation signal vector $\mathbf{D}$ [†]. Though the SNR performance of SICs is influenced by the error propagation, it is not easy to evaluate the error propagation theoretically. As has been done in the SNR performance analysis, we neglect the error propagation in the SNR performance analysis. Let $\underline{\mathbf{N}}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{T}}+N_{\mathrm{B}})\times 1}$ denote a detector input noise vector fed into the SIC, i.e., $\underline{\mathbf{N}}_{\mathrm{R}} = g_{\mathrm{R}}^{-1}\mathbf{Q}_{\mathrm{R}}^{\mathrm{H}}\overline{\mathbf{N}}$, the correlation matrix of the detector input noise vector is derived in (25). σ_{d}^2 and σ^2 indicate the power of the modulation signal and that of the AWGN. The SNR of the mth stream $\rho_{\mathrm{R}}(m) \in \mathbb{R}$ can be defined as,

$$\rho_{\mathrm{R}}(m) = \frac{\gamma_{\mathrm{R}}^2(m)\,\sigma_{\mathrm{d}}^2}{\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{R}}\,\underline{\mathbf{N}}_{\mathrm{R}}^{\mathrm{H}}\right]_{m,m}}. \tag{26}$$

3.3.2 THP Based on MMSE

Even when precoding based on the MMSE is applied to the system, similar as the system with the linear precoding, a detector input vector can be obtained by multiplying the received signal vector with inverse of the normalization factor g_{T}^{-1} and the transform matrix $\mathbf{S}_{\mathrm{T}}$ as,

$$\mathbf{Z}_{\mathrm{T}} = g_{\mathrm{T}}^{-1}\mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\overline{\mathbf{Y}} = \dot{\mathbf{D}} + g_{\mathrm{T}}^{-1}\mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\overline{\mathbf{N}}. \tag{27}$$

As is seen, the region detection can be used to estimate the modulation signal vector. Let a detector input noise vector $\underline{\mathbf{N}}_{\mathrm{T}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times 1}$ be defined as $\underline{\mathbf{N}}_{\mathrm{T}} = g_{\mathrm{T}}^{-1}\mathbf{S}_{\mathrm{T}}^{\mathrm{H}}\overline{\mathbf{N}}$, the correlation matrix of the detector input noise vector is shown in (28).

When the feedback filter output signals are assumed to be uniformly distributed, all the diagonal elements of the correlation matrix Φ_{V} are reduced to $\frac{M^2}{6}$. If the elements of the feedback filter output vector are uncorrelated from each other, the matrix Φ_{V} approximately results in the diagonal matrix $\frac{M^2}{6}\mathbf{I}_{(N_{\mathrm{A}}+N_{\mathrm{R}})\times(N_{\mathrm{A}}+N_{\mathrm{R}})}$ [††], which is used in the derivation of (28). The SNR of the mth stream $\rho_{\mathrm{T}}(m)$ can be defined as,

$$\rho_{\mathrm{T}}(m) = \frac{\sigma_{\mathrm{d}}^2}{\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{T}}\,\underline{\mathbf{N}}_{\mathrm{T}}^{\mathrm{H}}\right]_{m,m}}. \tag{29}$$

3.4 Discussion

As is shown in (25) and (26) as well as (28) and (29), the SNR depends on the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$. Because this is a first step of the research, we would like to start with the simplest example in order to grasp the basic characteristics of the proposed overloaded MIMO spatial multiplexing[†††]. We would like to apply diagonal matrices and null matrices to those matrices in the following sections, as the simplest example.

3.5 Linear Precoding

As is shown in (25), the matrix $\mathbf{J}_1$ should be set to the null matrix in order to reduce the noise power of the 1st to N_{R}th streams. Besides, $\mathbf{J}_2 = \mathbf{J}_3 = \sqrt{\gamma_{\mathrm{R}}}\mathbf{I}_{N_{\mathrm{T}}}$ looks the simplest where $\gamma_{\mathrm{R}} \in \mathbb{R}$ represents a constant. However, this setting constrains the number of the spatially multiplexed signal streams to $2N_{\mathrm{T}}$, because this setting imposes the parameters to satisfy $N_{\mathrm{A}} = N_{\mathrm{B}} = N_{\mathrm{T}}$. In other words, this setting gets rid of the freedom to set the number of the spatially

[†]The transform matrix $\mathbf{Q}_{\mathrm{R}}$ transforms the extended channel matrix into the triangular matrix $\mathbf{R}_{\mathrm{R}}$, which makes it possible to apply the SIC at the receiver.

[††]The correlation matrix is given with the assumption that all the streams are added with the Gaussian integer multiples. The derivation on the assumption has been shown in [40].

[†††]The optimization of the those matrices is definitely one of our important future works.

multiplexed signal streams. On the other hand, let γ_{R} be set as $\gamma_{\mathrm{R}} = \frac{\sigma^2}{\sigma_{\mathrm{d}}^2 g^2} = \frac{N_{\mathrm{T}}\sigma^2}{P_0}$, it has the probability that the following setting can equalize the noise power of all the streams.

$$\mathbf{J}_1 = \mathbf{0}_{N_{\mathrm{R}}\times N_{\mathrm{B}}}, \quad (\mathbf{J}_2 \quad \mathbf{J}_3) = \sqrt{\gamma_{\mathrm{R}}}\mathbf{I}_{N_{\mathrm{A}}} \tag{30}$$

When the setting written in (30) is used, the parameter N_{B} is expressed with the other parameters, i.e., $N_{\mathrm{B}} = N_{\mathrm{A}} - N_{\mathrm{T}}$. The number of the spatially multiplexed signal streams N_{S} gets equal to N_{A}. In a word, $N_{\mathrm{S}} = N_{\mathrm{A}}$.

When the above matrices are applied to the extended channel matrix, an extended channel matrix $\overline{\mathbf{H}}_{\mathrm{R}} \in \mathbb{C}^{(N_{\mathrm{R}}+N_{\mathrm{A}})\times N_{\mathrm{A}}}$ for the linear precoding is expressed as[†],

$$\overline{\mathbf{H}}_{\mathrm{R}} = \begin{pmatrix} \mathbf{H} & \mathbf{0}_{N_{\mathrm{R}}\times(N_{\mathrm{A}}-N_{\mathrm{T}})} \\ \sqrt{\gamma_{\mathrm{R}}}\mathbf{I}_{N_{\mathrm{A}}\times N_{\mathrm{A}}} \end{pmatrix} \tag{31}$$

When the above setting is applied, if the transform matrix is unitary, i.e., $\mathbf{S}_{\mathrm{R}}^{\mathrm{H}}\mathbf{S}_{\mathrm{R}} = \mathbf{I}_{(N_{\mathrm{T}}+N_{\mathrm{B}})\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$, the correlation in the detector input noise vector is reduced to the following equation.

$$\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{R}}\,\underline{\mathbf{N}}_{\mathrm{R}}^{\mathrm{H}}\right] = \frac{N_{\mathrm{T}}\sigma_{\mathrm{d}}^2}{P_0}\sigma^2\mathbf{I}_{N_{\mathrm{A}}} \tag{32}$$

As is described above, all the streams get uncorrelated and the power of them is equalized. By substituting the noise power in (32) for (26), the SNR of the mth stream $\rho_{\mathrm{R}}(m)$ is rewritten as,

$$\rho_{\mathrm{R}}(m) = \frac{\gamma_{\mathrm{R}}^2(m)}{N_{\mathrm{T}}}\frac{P_0}{\sigma^2}. \tag{33}$$

While the SNR is described in (33) when the transform matrix $\mathbf{S}_{\mathrm{R}}$ is unitary, the SNR is dependent on not only the diagonal elements $\gamma_{\mathrm{R}}^2(m)$ and the noise power but also the characteristics of the transform matrix, when the transform matrix is not unitary. Therefore, the SNR performance is evaluated by computer simulation in Sect. 4.

However, the above discussion assumes that $N_{\mathrm{B}}\,(= N_{\mathrm{A}} - N_{\mathrm{T}})$ is non-negative, i.e., $N_{\mathrm{A}} \geq N_{\mathrm{T}}$. If the N_{A} is set to be smaller than N_{T}, i.e., $N_{\mathrm{A}} < N_{\mathrm{T}}$, N_{B} gets to be zero, $N_{\mathrm{B}} = 0$, and the number of the spatially multiplexed signal streams is fixed to N_{T}. In fact, the matrices $\mathbf{J}_1$ and $\mathbf{J}_3$ become null, and the matrix $\mathbf{J}_2$ is reduced as,

$$\mathbf{J}_1 = \mathbf{J}_3 = \varnothing, \quad \mathbf{J}_2 = \sqrt{\gamma_{\mathrm{R}}}\left(\mathbf{I}_{N_{\mathrm{A}}} \quad \mathbf{0}_{N_{\mathrm{A}}\times(N_{\mathrm{T}}-N_{\mathrm{A}})}\right). \tag{34}$$

3.6 THP Based on MMSE

We would like to apply the above successful discussion in this section for the THP. However, the settings described in the previous section can not be directly applied to the THP based on the MMSE, because the setting does not meet the requirement in (10). As is described in (9) and (10), the Hermite transpose of the extended channel matrix should be slim to apply it to the THP. Since the extended channel matrix in (31) is slim, it is a good start point that the matrix is used as an Hermite transpose of the extended channel matrix for the THP. However, the channel matrix $\mathbf{H}$ needs to be replaced with its Hermite transpose. In addition, $\mathbf{J}_1$ and $\mathbf{J}_2$ are renamed as $\mathbf{J}_2$ and $\mathbf{J}_1$ in the matrix to be content with the definition of the extended channel matrix. If the setting is applied, the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ are written as,

$$\mathbf{J}_2 = \mathbf{0}_{N_{\mathrm{A}}\times N_{\mathrm{T}}}, \quad \left(\mathbf{J}_1^{\mathrm{T}} \quad \mathbf{J}_3^{\mathrm{T}}\right)^{\mathrm{T}} = \sqrt{\gamma_{\mathrm{T}}}\mathbf{I}_{N_{\mathrm{B}}} \tag{35}$$

In (35), the sizes of the channel matrices are adjusted to be consistent with the definition of the extended channel matrix, i.e., $N_{\mathrm{B}} = N_{\mathrm{A}} + N_{\mathrm{R}}$. In a word, the number of the spatially multiplexed signal streams N_{S} becomes equal to $N_{\mathrm{B}}\,(= N_{\mathrm{A}} + N_{\mathrm{R}})$.

If the above replacement is used, consequently, the extended channel matrix $\mathbf{H}_{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{B}}\times(N_{\mathrm{T}}+N_{\mathrm{B}})}$ for the THP based on the MMSE can be derived as,

$$\overline{\mathbf{H}}_{\mathrm{T}}^{\mathrm{H}} = \begin{pmatrix} \mathbf{H}^{\mathrm{H}} & \mathbf{J}_2^{\mathrm{H}} \\ \mathbf{J}_1^{\mathrm{H}} & \mathbf{J}_3^{\mathrm{H}} \end{pmatrix} = \begin{pmatrix} \mathbf{H}^{\mathrm{H}} & \mathbf{0}_{N_{\mathrm{T}}\times(N_{\mathrm{B}}-N_{\mathrm{R}})} \\ \sqrt{\gamma_{\mathrm{T}}}\mathbf{I}_{N_{\mathrm{B}}\times N_{\mathrm{B}}} \end{pmatrix}. \tag{36}$$

Because the parameter setting $\gamma_{\mathrm{T}} = \frac{N_{\mathrm{R}}\sigma^2}{P_0}$ optimizes the transmission performance of THPs [40], we borrow the parameter setting for the proposed THP based on the MMSE[††]. Let the transform matrix $\mathbf{S}_{\mathrm{T}}$ be decomposed as $\mathbf{S}_{\mathrm{T}} = \left(\mathbf{S}_{\mathrm{T},1}^{\mathrm{T}}, \mathbf{S}_{\mathrm{T},2}^{\mathrm{T}}\right)^{\mathrm{T}}$ where $\mathbf{S}_{\mathrm{T},1} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{B}}}$ and $\mathbf{S}_{\mathrm{T},2} \in \mathbb{C}^{N_{\mathrm{A}}\times N_{\mathrm{B}}}$ represent submatrices of the transform matrix, as a result, the noise correlation is reduced to,

$$\begin{aligned}\mathrm{E}\left[\underline{\mathbf{N}}_{\mathrm{T}}\,\underline{\mathbf{N}}_{\mathrm{T}}^{\mathrm{H}}\right] = \frac{\sigma^2}{g^2}\Big\{&\mathbf{S}_{\mathrm{T},1}^{\mathrm{H}}\mathbf{S}_{\mathrm{T},1} + \Big(\mathbf{S}_{\mathrm{T},1}^{\mathrm{H}}\mathbf{Q}_{\mathrm{T},2,1}\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\mathbf{Q}_{\mathrm{T},2,1}^{\mathrm{H}}\mathbf{S}_{\mathrm{T},1} \\ &+ \mathbf{S}_{\mathrm{T},2}^{\mathrm{H}}\mathbf{Q}_{\mathrm{T},2,2}\boldsymbol{\Gamma}_{\mathrm{T}}^{-2}\mathbf{Q}_{\mathrm{T},2,2}^{\mathrm{H}}\mathbf{S}_{\mathrm{T},2}\Big)\Big\}.\end{aligned} \tag{37}$$

In (37), $\mathbf{Q}_{\mathrm{T},2,1} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{B}}}$ and $\mathbf{Q}_{\mathrm{T},2,2} \in \mathbb{C}^{N_{\mathrm{A}}\times N_{\mathrm{B}}}$ denote submatrices of the matrix $\mathbf{Q}_{\mathrm{T},2}$, which are defined as $\mathbf{Q}_{\mathrm{T},2} = \left(\mathbf{Q}_{\mathrm{T},2,1}^{\mathrm{T}} \quad \mathbf{Q}_{\mathrm{T},2,2}^{\mathrm{T}}\right)^{\mathrm{T}}$. Nevertheless, the above analysis is available if $N_{\mathrm{A}}\,(= N_{\mathrm{B}} - N_{\mathrm{R}})$ is non-negative, i.e., $N_{\mathrm{B}} \geq N_{\mathrm{R}}$.

When N_{B} is set to be less than N_{R}, i.e., $N_{\mathrm{B}} < N_{\mathrm{R}}$, N_{A} is imposed to be set to zero, $N_{\mathrm{A}} = 0$, and the number of the spatially multiplexed signal streams is fixed to N_{R}. As a result, the matrices $\mathbf{J}_2$ and $\mathbf{J}_3$ become null, and the matrix $\mathbf{J}_1$ is reduced as,

$$\mathbf{J}_2 = \mathbf{J}_3 = \varnothing, \quad \mathbf{J}_1 = \sqrt{\gamma_{\mathrm{T}}}\left(\mathbf{I}_{N_{\mathrm{B}}} \quad \mathbf{0}_{N_{\mathrm{B}}\times(N_{\mathrm{R}}-N_{\mathrm{B}})}\right)^{\mathrm{T}}. \tag{38}$$

The SNR of the m stream $\rho_{\mathrm{T}}(m)$ is rewritten by substituting the noise variance in (37) for (29). Compared with the

[†]The extended channel matrix $\overline{\mathbf{H}}_{\mathrm{R}}$ is identical to the matrix $\overline{\mathbf{H}}$. However, because the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ are designed suitable for the linear precoding, the extended matrix dares to be named as $\overline{\mathbf{H}}_{\mathrm{R}}$.

[††]The optimization of the parameter γ_{T} for the proposed THP based on the MMSE is one of our future works.

SNR performance of the linear precoding, that of the THP is a little bit more complex.

3.7 Characteristics of Proposed Scheme and Other Issues

As is described in Sect. 1, many overloaded MIMO techniques have been proposed before. Some of them achieve fair transmission performance with relatively high computational complexity. For instance, the overloaded MIMO with the virtual channels needs sequence estimation like the maximum likelihood sequence estimation (MLSE) to achieve such performance [30]. Another overloaded MIMO with the virtual channels requires a complex non-linear optimization every packet [31]. In those techniques, the number of the spatially multiplexed signal streams is limited by that of the antennas on the transmitter. On the other hand, the proposed overloaded MIMO spatial multiplexing can raise the number of the spatially multiplexed signal streams to more than that of the antennas on the receiver or the transmitter, which characterizes the proposed multiplexing. Because such overloaded spatial multiplexing can not be implemented other than the proposed overloaded MIMO, it is impossible to compare the proposed overloaded MIMO spatial multiplexing with the conventional techniques in terms of the transmission performance and the computational complexity.

To implement them, one of the overloaded MIMO systems with the virtual channels optimizes the precoding weights and feeds them back to the precoders at the transmitter, every slot [31]. While no information needs to be fed back in the MIMO system proposed in the paper [30], the packet length should be small if the computational complexity is taken into account. On the other hand, the proposed overloaded MIMO spatial multiplexing should share the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ among all the receiver and the transmitter before the communication starts. However, the matrices are products of the identity matrices and the scalar value as defined in (30), (34), (35), and (38). While the scalar value needs to be set based on the noise variance, those identity matrices can be set in the transmitter and the receiver when they are configured at factories. If the identity matrices are not set at factories, it is enough to broadcast them for the receiver and the transmitter once when they are turned on. In a word, the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$ don't need to be frequently exchanged between the transmitter and the receiver. The noise variance has to be sent not only in the proposed multiplexing but also in systems with the MMSE precoding. The amount of the information to send the noise variance is negligible small compared with that of the information bits to send. This information exchange before the communication does not cause any serious problem.

On the other hand, the proposed overloaded MIMO spatial multiplexing needs the spatial filters at the receiver as shown in (24) and (27). The need for the spatial filters at the receiver restricts the proposed overloaded MIMO spatial multiplexing to single-user MIMO systems.

4. Computer Simulation

The performance of the proposed overloaded MIMO spatial multiplexing is evaluated by computer simulation. The modulation scheme is quaternary phase shift keying (QPSK) [41]. The proposed spatial multiplexing enables the number of the spatially multiplexed signal streams to be greater than the maximum number of the antennas in the system, i.e.,$\max[N_\mathrm{T}, N_\mathrm{R}]$. This means that the number of the spatially multiplexed signal streams exceeds that of the eigenvalues in the channel. To confirm how the number of the eigenvalues affects the transmission performance, we evaluate the transmission performance in not only independent and identically distributed (IID) Rayleigh fading channels but also Keyhole channels based on the Jakes' model [42]. The number of the transmit antennas N_T and that of the received antennas N_R are all set to 2. The number of the spatially multiplexed signal streams N_S is increased to 4 or 6, which correspond with the overloading ratio of 2.0 or 3.0, respectively. The LLL algorithm is used for the lattice reduction. The simulation parameters are listed in Table 1. The following performance comparison is performed on the same assumption, for instance, the error correction coding is not applied as listed in the Table. Only the QR-decomposition makes the difference in the performance as shown in the following sections.

4.1 SNR Performance

The signal power to noise power ratio (SNR) distribution of the proposed overloaded spatial multiplexing is analyzed to grasp the transmission performance. The number of the spatially multiplexed signal streams N_S is set to 4 in Sect. 4.1. This subsection shows some figures where the abscissa is the SNR in dB and the ordinate is the cumulative distribution function. The E_b/N_0 is set to 20 dB in this section. While the SNR performance are analyzed above theoretically, the SNR performance is evaluated by means of computer simulations in this section. Since the detector input signals are defined in (24) and (27), the SNR measurement depends on the precoding schemes. When the linear precoding is employed, the noise power can be measured as $\sigma_\mathrm{R}^2(m) = \mathrm{E}\left[\left|\frac{1}{\gamma_\mathrm{R}} z_\mathrm{R}(m) - d(m) - \sum_{k=N_\mathrm{T}}^{m-1} r_\mathrm{R}(m,k)\,\bar{d}(k)\right|^2\right]$ where $\sigma_\mathrm{T}^2(m)$, $z_\mathrm{T}(m)$, $d(m)$, and $\bar{d}(k)$ denote a noise power in an mth detector input signal, an mth entry of the detector input signal $\mathbf{Z}_\mathrm{T}$, that of the modulation signal vector $\mathbf{D}$, and a kth estimated modulation signal, respectively. When the

Table 1 Simulation parameters.

Modulation	QPSK
Channel model	IID, Keyhole based on Rayleigh fading
$(N_\mathrm{T}, N_\mathrm{R})$	(2, 2)
Channel estimation	Perfect
Lattice reduction	LLL algorithm
Overloading Ratio	2.0, 3.0
Error correction coding	N.A.

THP is used. on the other hand, the noise power can be measured as $\sigma_{\mathrm{T}}^2(m) = \mathrm{E}\left[|z_{\mathrm{T}}(m) - d(m)|^2\right]$ where $\sigma_{\mathrm{T}}^2(m)$ and $z_{\mathrm{T}}(m)$ represent a noise power in the mth detector input signal and an mth entry of the detector input signal $\mathbf{Z}_{\mathrm{T}}$, respectively. Hence, the SNR $\rho_\Omega(m)$ can be calculated as $\rho_\Omega(m) = \frac{\sigma_{\mathrm{d}}^2}{\sigma_\Omega^2(m)}$ where Ω takes R or T.

Figure 2 and Fig. 3 show the distribution functions of the SNR of all the streams when the linear precoding and the THP with the SQRD are applied to the proposed multiplexing, respectively. While the probability that the SNR distribution of 3rd and 4th streams gets less than 5 dB is about 10^{-3}, the SNR of the 1st and the 2nd streams is fixed about 0 dB. The performance is greatly dependent on the distribution of the diagonal elements of the matrix $\mathbf{\Gamma}_\Omega \quad \Omega = \mathrm{R}$ or T. While the number of the eigenvalues in the channel is equal to 2 in the setting of the section, the number of the eigenvalues in the extended channel matrices is the same to that of the signal streams, i.e., 4 in spite of the type of the precoding. Also, the diagonal matrix $\mathbf{\Gamma}_\Omega$ has 4 diagonal non-zero elements. On the other hand, when the E_{b}/N_0 is high enough, $\sqrt{\gamma_\Omega}$ is smaller than all the eigenvalues in the channel matrix $\mathbf{H}$. Even if the SQRD is applied, the eigenvalues are concentrated into only $2\,(= N_{\mathrm{R}} = N_{\mathrm{T}})$ diagonal elements, while the other diagonal elements are almost same to $\sqrt{\gamma_\Omega}$, which causes that the SNR of the two streams gets much less than that of the others. This causes half of the detector input signals to be almost zero, while the other input signals are non-zero values. In addition, the non-diagonal elements in the 1st and 2nd row of the upper triangular matrix $\mathbf{R}_{\mathrm{R}}$ become zero[†]. This causes the noise power $\sigma_{\mathrm{R}}(m)$ rewritten as $\sigma_{\mathrm{R}}^2(m) = \mathrm{E}\left[\left|\frac{1}{\gamma_{\mathrm{R}}} z_{\mathrm{R}}(m) - d(m)\right|^2\right]$ for $m = 1, 2$. Hence, the noise power can be reduced to $\sigma_{\mathrm{R}}(m) = \mathrm{E}[|d(m)|^2] = \sigma_{\mathrm{d}}^2$. Eventually, the SNR of the 1st and the 2nd streams is reduced to 0 dB.

Even when the THP is applied, the similar performance can be seen due to the reason described above, as long as the SQRD is applied. Therefore, the SNR of the 1st and the 2nd streams is fixed to 0 dB.

Figure 4 and Fig. 5 show the distribution functions of the SNR of all the streams when the linear precoding and the THP with the lattice reduction are applied to the proposed multiplexing, respectively. Similar as the performance with the SQRD, the SNR of the 1st and the 2nd is about 0 dB, while that of the 3rd and the 4th streams is much bigger than 0 dB. Even when the lattice reduction is applied for QR-decomposition, the 1st and 2nd diagonal elements of the diagonal matrix $\mathbf{\Gamma}_\Omega$ is much smaller than the 3rd and 4th diagonal elements, which causes the SNR performance to degrade as shown in Fig. 4 when the linear precoding is applied. Even if the THP is used, the similar SNR performance can be obtained as shown in Fig. 5.

Figure 6 and Fig. 7 show the SNR distribution functions

[†]The theoretical performance analysis of those phenomena is one of our future works.

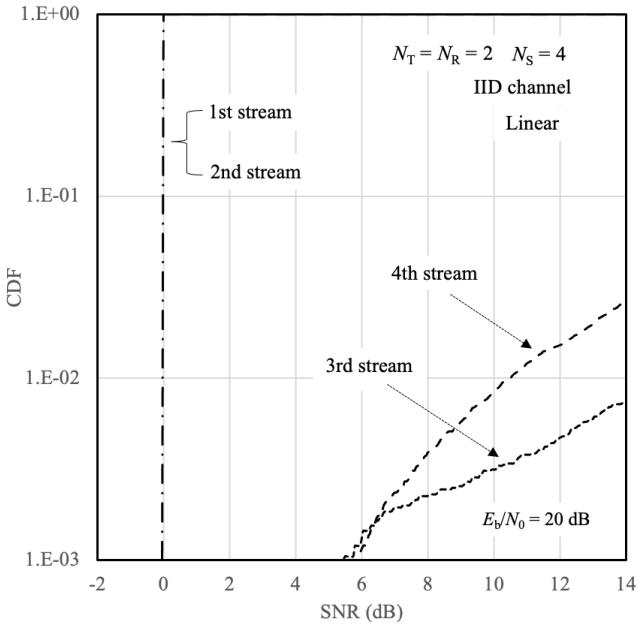


Fig. 2 SNR distribution of linear precoding with SQRD.

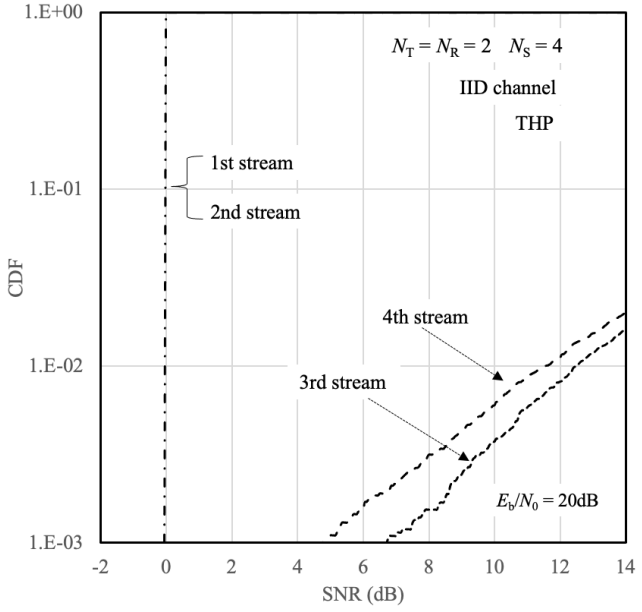


Fig. 3 SNR distribution of THP with SQRD.

when the linear precoding and the THP with the equal gain transform are applied to the proposed multiplexing, respectively. "Theory" in Fig. 6 shows the theoretical performance written in (33)[††]. The theoretical SNR distributions of all the streams are same when the linear precoding is used. Since the equal gain transform equalizes the diagonal elements of the diagonal matrix $\mathbf{\Gamma}_\Omega$, the SNR of all the streams is equal-

[††]While the performance can be regarded as an upper bound, we dare to name the performance as "Theory", because we intend to emphasize that the performance is derived theoretically, and to show the performance in comparison with the simulation results.

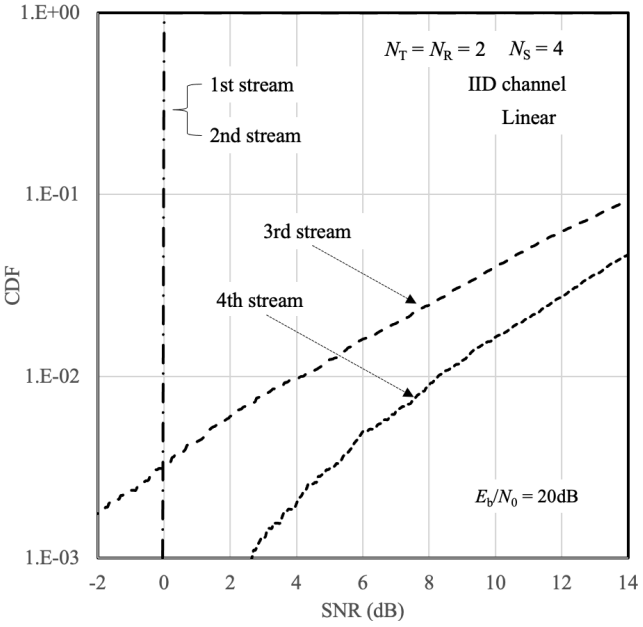


Fig. 4 SNR distribution of linear precoding with lattice reduction.

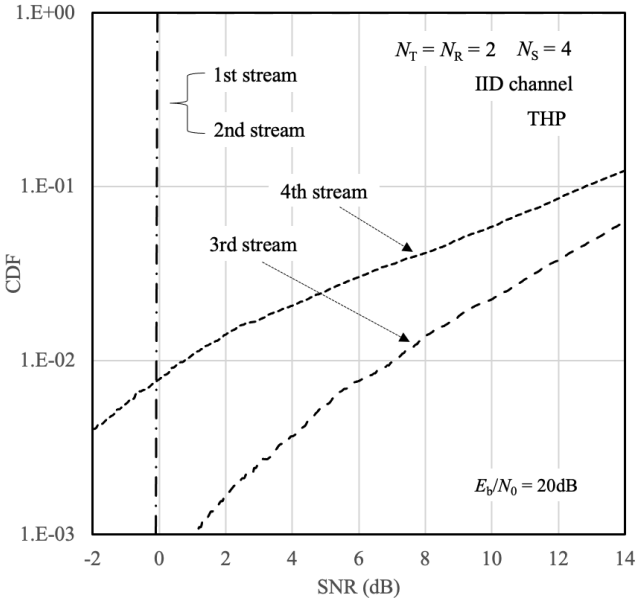


Fig. 5 SNR distribution of THP with lattice reduction.

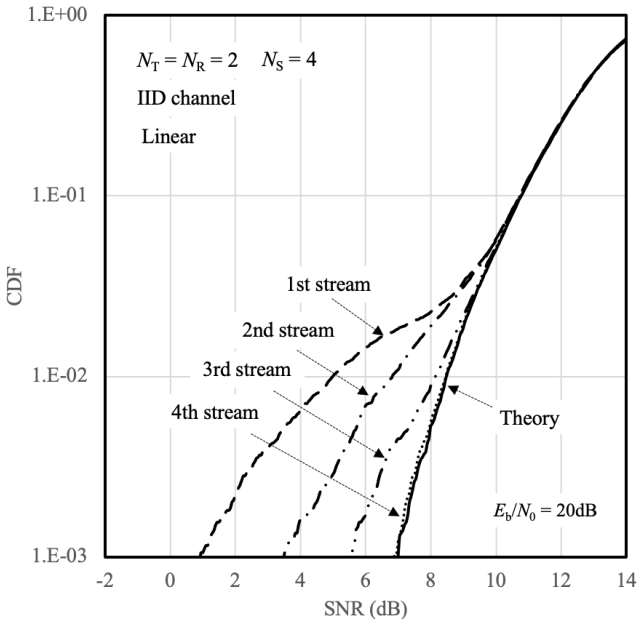


Fig. 6 SNR distribution of linear precoding with equal gain transform.

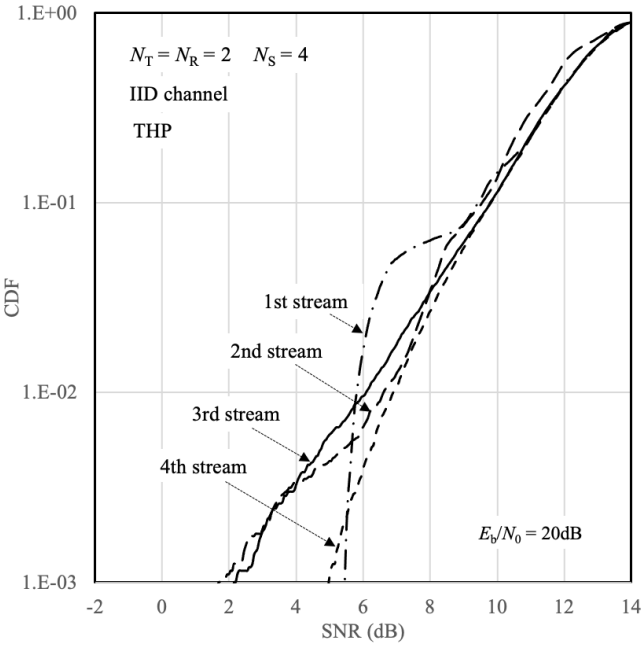


Fig. 7 SNR distribution of THP with equal gain transform.

ized, which agrees with the simulation result shown in Fig. 6. Actually, the SNR distribution of the signal streams is different from each other because of the error propagation. While the SNR distribution of the 4th stream is the worst, that of the 1st stream is the best, which is the same to the theoretical SNR performance, because the propagation error does not happen in the first stream. Even when the THP is employed, the similar performance is expected, because of the performance of the equal gain transform. In fact, the SNR distributions of all the streams are only a little bit deviated as shown in Fig. 7. As is shown in the figure, the probability that the SNR is less than 10 dB is about 10^{-1}. The SNR is called a required SNR for 10% outage probability. On the other hand, when the linear precoding is applied, the required SNR for 10% outage probability is improved to about 11 dB as shown in Fig. 6. In a word, the linear precoding achieves about 1.0 dB better SNR performance than the THP at the 10% outage probability.

4.2 BER Performance

Since the three QR-decomposition techniques are consid-

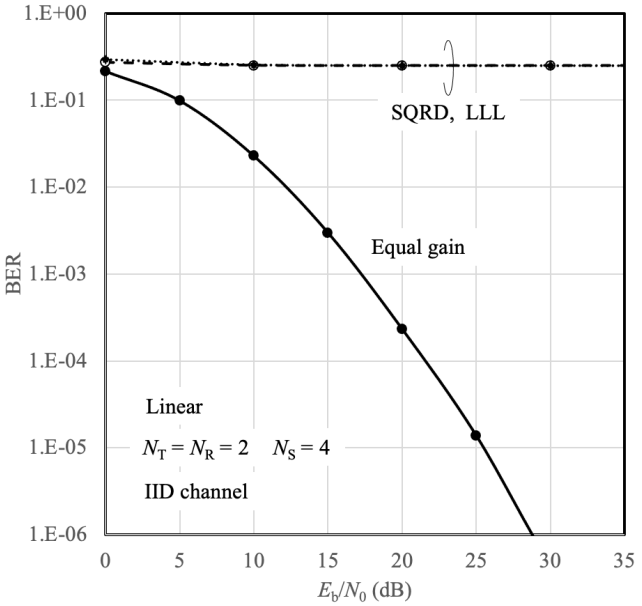


Fig. 8 BER Performance of Linear Precoding with QR-decomposition techniques.

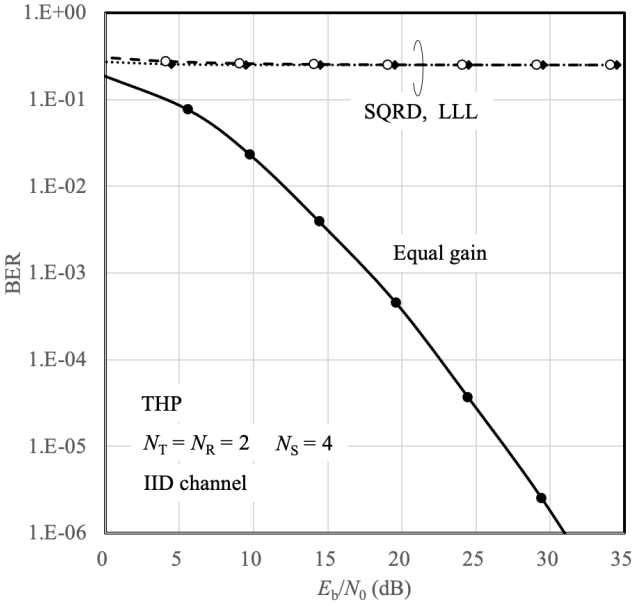


Fig. 9 BER Performance of THP with QR-decomposition techniques.

ered, those techniques are compared in terms of the BER performance in this section. The number of the spatially multiplexed signal streams is fixed to 4, which means that the overloading ratio is 2.0. The channel model is the IID based on the Jake's model.

4.2.1 Performance VS. QR-Decomposition Techniques

Figure 8 shows the BER performances of the proposed overloaded MIMO spatial multiplexing when the linear precoding is applied. In the figure, the three QR-decomposition techniques are compared in terms of the BER. When the LLL and the SQRD are used, the proposed spatial multiplexing can't reduce the irreducible error to less than 0.3, which agrees with the SNR distributions shown in Fig. 2 and Fig. 4. On the other hand, the equal gain transform makes the proposed spatial multiplexing achieve superior performance, which also agrees with the SNR distribution shown in Fig. 6.

Figure 9 shows the BER performance when the THP is used. Also, the three QR-decomposition techniques are employed. Similar as Fig. 8, the irreducible error appears at about 0.3 when the LLL and the SQRD are used for the QR-decomposition. On the other hand, the equal gain transform enables the proposed overloaded MIMO spatial multiplexing to attain the superior performance. Besides, the linear precoding achieves about 1.0 dB better BER performance than the THP at the BER of 10^{-5}, which agrees with the SNR performance comparison described at the end of Sect. 4.1.

4.2.2 BER Performance in Keyhole Channel

As is shown in the previous sections, the proposed overloaded MIMO spatial multiplexing achieves superior performance, even though the number of the spatially multiplexed signal streams is twice as many as that of the eigenvalues in the channel. This section evaluates the BER performance in the channels with less than 2 eigenvalues. We apply a keyhole channel for the performance evaluation. Let $\mathbf{h}_R \in \mathbb{C}^{N_R \times 1}$ and $\mathbf{h}_T \in \mathbb{C}^{1 \times N_T}$ denote vectors defined as $\mathbf{h}_R = (h_R(1) \cdots h_R(N_R))^T$ and $\mathbf{h}_T = (h_T(1) \cdots h_T(N_T))$ where $h_\Omega(m) \in \mathbb{C}$ represents a channel gain, Ω = R or T, the keyhole channel is defined as $\mathbf{H} = \mathbf{h}_R \mathbf{h}_T$. In the performance evaluation, the channel gains are generated based on the Jake's model for fair performance comparison. While this keyhole channel is a kind of Rayleigh fading channels, the number of the eigenvalues is reduced to 1.

Figure 10 shows the BER performance of the proposed overloaded MIMO spatial multiplexing in the Keyhole channel. In the figure, the equal gain transform is employed for the QR-decomposition. The performances of not only the linear precoding but also the THP are illustrated in the figure. The performance in the IID channel is added as a reference in the figure. Whereas the performance in the keyhole channel is much worse than that in the IID channel, however, the irreducible error does not appear up to the BER of 10^{-3}, This means that the decrease in the number of eigenvalues in channels does not cause the irreducible error up to 10^{-3}, even though the performance is degraded as the number of the eigenvalues in channels is decreased. The linear precoding is able to make the proposed overloaded MIMO spatial multiplexing achieve better BER performance than the THP even in the keyhole channel.

4.3 Overloading Ratio

Figure 11 compares the BER performance with the overloading ratio of 2.0 and that of 3.0 in the IID channel. In

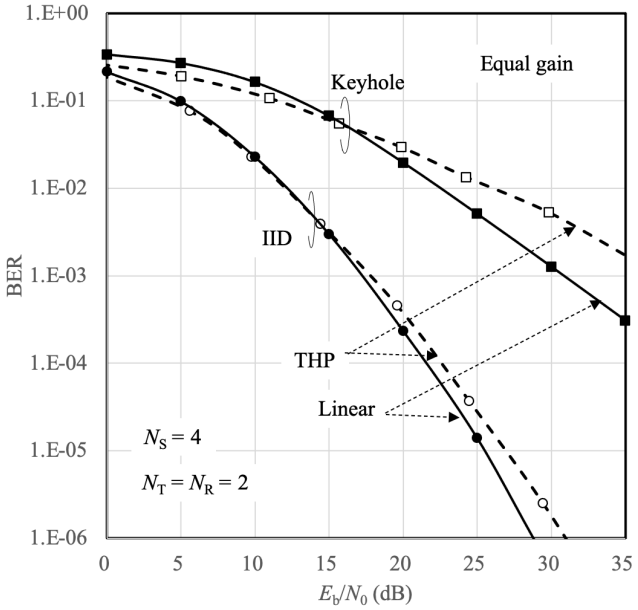


Fig. 10 BER performance in keyhole channels.

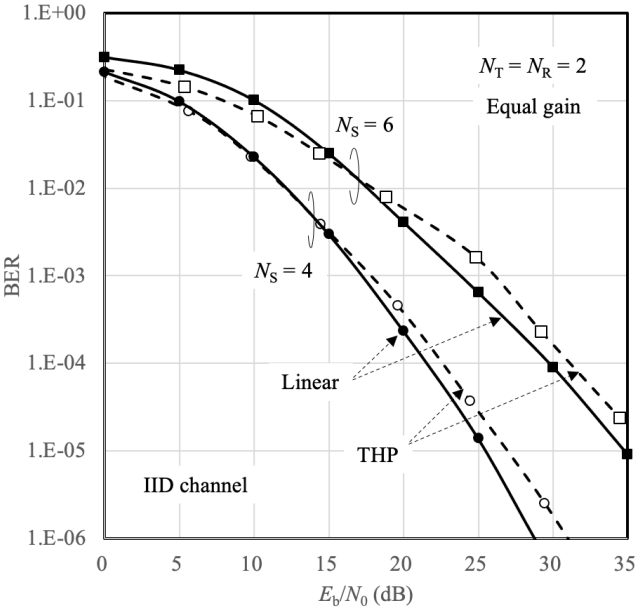


Fig. 11 BER performance of linear precoding in overloaded MIMO channels.

the figure, the equal gain transform is used. The number of the receive antennas and that of the transmit antennas are all fixed to 2. The overloading ratio is raised by changing the extended channel matrices. For example, the overloading ratios of 2.0 and 3.0 are implemented with setting the parameter N_A to 4 and 6, respectively, when the linear precoding is applied. When the THP is used, the parameter N_B is also set to 4 and 6 for the overloading ratio of 2.0 and 3.0, respectively. The proposed overloaded MIMO spatial multiplexing with the THP outperforms that with the linear filter in the E_b/N_0 of less than 15 dB, while the performance of the proposed spatial multiplexing with the THP is inferior to that with the linear filter in the other E_b/N_0 region[†]. The proposed overloaded MIMO spatial multiplexing achieves superior transmission performance even if the overloading ratio is raised to 3.0. The proposed overloaded MIMO spatial multiplexing with the linear precoding achieves about 1.0 dB better BER performance than that with the THP, despite of the overloading ratio.

[†]Such performances probably depends on the the choice of the matrices $\mathbf{J}_1$, $\mathbf{J}_2$, and $\mathbf{J}_3$. However, the performance analysis is one of our future works.

5. Conclusion

This paper has proposed overloaded MIMO spatial multiplexing that can increase the number of spatially multiplexed signal streams in spite of antenna settings on a receiver or a transmitter. The extension of a channel matrix with some matrices has been proposed to implement such overloaded signal transmission. Three tapes of QR-decomposition techniques such as the SQRD, the lattice reduction with the LLL algorithm, and the equal gain transform have been considered for precoding in the proposed overloaded MIMO spatial multiplexing. We apply linear precoding and the THP for the precoding, which can be used complementary. This paper shows some examples for the extended channel matrices and analyze the performance theoretically. The performance is confirmed by computer simulation. The linear precoding achieves better performance that the THP in the system with the example of the extended channel matrices. The equal gain transform makes the proposed overloaded MIMO spatial multiplexing achieve the best transmission performance in the three types of the QR-decomposition techniques. The proposed overloaded spatial multiplexing with the linear precoding based on the equal gain transform achieves 6 spatially multiplexed signal streams transmission with superior transmission performance, even though only two antennas are employed at both the receiver and the transmitter. In a word, the proposed overloaded MIMO spatial multiplexing is able to increase the number of the spatially multiplexed signal streams despite of the number of the antennas on both the receiver and the transmitter.

Acknowledgments

The work has been supported by JSPS KAKENHI JP21K04061 and 24K07475, the support center for advanced telecommunications technology research (SCAT), and Softbank Co. ltd.

References

[1] G.J. Foschini and M.J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," Wireless Pers. Commun., vol.6, no.3, pp.311–335, 1998.
[2] I.E. Telater, "Capacity of multi-antenna Gaussian channels," European Trans. Telecommun., vol.10, no.6, pp.585–595, 1999.
[3] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road

to large-scale MIMOs," IEEE Commun. Surveys Tuts., vol.17, no.4, pp.1941–1988, Fourthquarter 2015.
[4] G.J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multiple antennas," Bell Lab. Tech. J., vol.1, no.2, pp.41–59, 1996.
[5] P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," Proc. IEEE ISSSE-98, Pisa, Italy, Sept. 1998.
[6] Q.H. Spencer, A.L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," IEEE Trans. Signal Process., vol.52, no.2, pp.461–471, 2004.
[7] T. Abe, S. Tomisato, and T. Matsumoto, "A MIMO turbo equalizer for frequency-selective channels with unknown interference," IEEE Trans. Veh. Technol., vol.53, no.3, pp.476–482, 2003.
[8] E.G. Larsson, O. Edfors, F. Tufvesson, and T.L. Marzetta, "Massive MIMO for next generation wireless systems," IEEE Commun. Mag., vol.52, no.2, pp.186–195, Feb. 2014.
[9] M. Sakai, K. Kamohara, H. Iura, H. Nishimoto, K. Ishioka, Y. Murata, M. Yamamoto, A. Okazaki, N. Nonaka, S. Suyama, J. Mashino, A. Okamura, and Y. Okumura, "Experimental field trials on MU-MIMO transmissions for high SHF wide-band massive MIMO in 5G," IEEE Trans. Wireless Commun., vol.19, no.4, pp.2196–2207, April 2020.
[10] L. Lu, G.Y. Li, A.L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," IEEE J. Sel. Topics Signal Process., vol.8, no.5, pp.742–758, Oct. 2014.
[11] P. Som, T. Datta, A. Chockalingam, and B.S. Rajan, "Improved large-MIMO detection based on damped belief propagation," IEEE Information Theory Workshop on Information Theory (ITW), 2010.
[12] W. Fukuda, T. Abiko, T. Nishimura, T. Ohgane, Y. Ogawa, Y. Ohwatari, and Y. Kishiyama, "Low-complexity detection based on belief propagation in a massive MIMO system," IEEE 77th Veh. Technol. Conf. (VTC Spring), 2013.
[13] T. Takahashi, S. Ibi, and S. Sampei, "On normalization of matched filter belief in GaBP for large MIMO detection," IEEE 84th Veh. Technol. Conf. (VTC-Fall), 2016.
[14] R. Hoshyar, F.P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," IEEE Trans. Signal Process., vol.56, no.4, pp.1616–1626, April 2008.
[15] R. Stoica, G. Abreu, Z. Liu, T. Hara, and K. Ishibashi, "Massively concurrent non-orthogonal multiple access for 5G networks and beyond" IEEE Access, vol.7, pp.82080–82100, June 2019.
[16] S.M.A. Kazmi, N.H. Tran, T.M. Ho, D. Niyato, and C.S. Hong, "Coordinated device-to-device communication with non-orthogonal multiple access in future wireless cellular networks," IEEE Access, vol.6, pp.39860–39875, June 2018.
[17] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," IEICE Trans. Commun., vol.E98-B, no.3, pp.403–414, March 2015.
[18] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," Proc. IEEE, vol.105, no.12, pp.2347–2381, Dec. 2017.
[19] H. Yang, X. Fang, Y. Liu, X. Li, Y. Luo, and D. Chen, "Impact of overloading on link-level performance for sparse code multiple access," 25th Wireless and Optical Communication Conference (WOCC), 2016.
[20] M. Anan, M. Sawahashi, and Y. Kishiyama, "BLER performance of windowed-OFDM using faster-than-Nyquist signaling with 16QAM," 21st International Symposium on Wireless Personal Multimedia Communications (WPMC), 2018.
[21] K.K. Wong, A. Paulraj, and R.D. Murch, "Efficient high-performance decoding for overloaded MIMO antenna systems," IEEE Trans. Wireless Commun., vol.6, no.5, pp.1833–1843, May 2007.
[22] N. Surajudeen-Bakinde, X. Zhu, J. Gao, and A.K. Nandi, "Improved signal detection approach using genetic algorithm for overloaded MIMO systems," 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008.
[23] I. Shubhi and Y. Sanada, "Joint turbo decoding for overloaded MIMO-OFDM systems," IEEE Trans. Veh. Technol., vol.66, no.1, pp.433–442, Jan. 2017.
[24] X. Lian and D. Li, "On the application of sphere decoding algorithm in overload MIMO systems," IEEE International Conference on Information Theory and Information Security, Dec. 2010.
[25] S. Yoshikawa, S. Denno, and M. Morikura, "Complexity reduced lattice-reduction-aided MIMO receiver with virtual channel detection," IEICE Trans. Commun., vol.E96-B, no.1, 263–270, Jan. 2013.
[26] R. Hayakawa, K. Hayashi, and M. Kaneko, "Lattice reduction-aided detection for overloaded MIMO using slab decoding," IEICE Trans. Commun., vol.E99-B, no.8, pp.1697–1705, Aug. 2016.
[27] R. Hayakawa and K. Hayashi, "Convex optimization-based signal detection for massive overloaded MIMO systems," IEEE Trans. Wireless Commun., vol.16, no.11, pp.7080–7091, Nov. 2017.
[28] R. Shioji, T. Imammura, and, Y. Sanada, "Overloaded MIMO detection based on two-stage belief propagation with MMSE precancellation," IEEE Vehicular Technol. Conf. (VTC2021-Fall), Sept. 2021.
[29] T. Takahashi, S. Ibi, A. Tölli, and S. Sampei, "Subspace marginalized belief propagation for mmWave overloaded MIMO signal detection," IEEE Intern. Conf. Commun. (ICC2020), June 2020.
[30] D.K.C. So and Y. Lan, "Virtual receive antenna for overloaded MIMO layered space-time system," IEEE Trans. Commun., vol.60, no.6, pp.1610–1620, June 2012.
[31] G. Yang, Y. Zhou, and W. Xia, "Performance optimization for overloaded MIMO systems with virtual channel approach," Hindawi Wireless Commun. Mobile Comput., vol.2018, Article ID:9651378, pp.1–6, 2018.
[32] S. Denno, Y. Kawaguchi, H. Murata, and D. Umehara, "An iterative noise cancelling receiver with soft-output LR-aided detection for collaborative reception," 19th International Symposium on Wireless Personal Multimedia Communications (WPMC 2016), Shenzhen China, Nov. 2016.
[33] S. Denno, T. Inoue, T. Fujiwara, and Y. Hou, "Low complexity soft input decoding in an iterative linear receiver for overloaded MIMO," IEICE Trans. Commun., vol.E103-B, no.5, pp.600–608, May 2020.
[34] D. Wübben, R. Böhnke, V. Kühn, and K.-D. Kammeyer, "MMSE extension of V-BLAST based on sorted QR decomposition," IEEE 58th Veh. Technol. Conf. (VTC2003-Fall), Orlando Fl, USA, Oct. 2003.
[35] D. Wübben, D. Seethaler, J. Jalden, and G. Mats, "Lattice reduction," IEEE Signal Process. Mag., vol.28, no.3, pp.70–91, 2011.
[36] A.K. Lenstra, H.W. Lenstra, Jr., and L. Lovász, "Factoring polynomials with rational coefficients," Math. Ann., vol.261, no.4, pp.515–534, 1982.
[37] S. Denno, Y. Kawaguchi, T. Inoue, and Y. Hou, "A novel low complexity lattice reduction-aided iterative receiver for overloaded MIMO," IEICE Trans. Commun., vol.E102-B, no.5, pp.1045–1054, May 2019.
[38] J.-K. Zhang, A. Kavcic, and K.M. Wong, "Equal-diagonal QR decomposition and its application to precoder design for successive-cancellation detection," IEEE Trans. Inf. Theory, vol.51, no.1, pp.154–172, 2005.
[39] M. Joham, W. Utschick, and J.A. Nossek, "Linear transmit processing in MIMO communications systems," IEEE Trans. Signal Process., vol.53, no.8, pp.2700–2712, 2005.
[40] K. Kusume, M. Joham, W. Utschick, and G. Bauch, "Cholesky factorization with symmetric permutation applied to detecting and precoding spatially multiplexed data streams," IEEE Trans. Signal Process., vol.55, no.6, pp.3089–3103, 2007.
[41] J.G. Proakis and M. Salehi, Digital Communications, 5th ed., McGraw-Hill, 2008.
[42] W.C. Jakes, Microwave Mobile Communications, IEEE Press, 1994.

Satoshi Denno received the M.E. and Ph.D. degrees from Kyoto University, Kyoto, Japan in 1988 and 2000, respectively. He joined NTT radio communications systems labs, Yokosuka, Japan, in 1988. He was seconded to ATR adaptive communications research laboratories, Kyoto, Japan in 1997. From 2000 to 2002, he worked for NTT DOCOMO, Yokosuka, Japan. In 2002, he moved to DOCOMO communications laboratories Europe GmbH, Germany. From 2004 to 2011, he worked as an associate professor at Kyoto University. Since 2011, he is a full professor at Okayama University. From the beginning of his research career, he has been engaged in the research and development of digital mobile radio communications. In particular, he has considerable interests in channel equalization, array signal processing, Space time codes, spatial multiplexing, and multimode reception. He won the Best paper award of the 19th international symposium on wireless personal multimedia communications (WPMC2016), and the outstanding paper award of the 23rd international conference on advanced communications technology (ICAT2021). He received the excellent paper award from the IEICE in 1995 and the best paper award from the IEICE communication society in 2020, respectively.

Takumi Sugimoto received B.S. and M.S. degrees from Okayama University, Japan, in 2022 and 2024, respectively. He joined with Murata Manufacturing Co. ltd., in 2024. His research interests include signal processing, wireless communication systems, and overloaded MIMO systems.

Koki Matoba received B.S. degree from Okayama University, Japan, in 2023. He is a master course student in Graduate School of Environmental, life, Natural Science and Technology, Okayama University. His research interests include signal processing, wireless communication systems, and overloaded MIMO systems.

Yafei Hou received his Ph.D. degrees from Fudan University, China and Kochi University of Technology (KUT), Japan in 2007. He was a post-doctoral research fellow at Ryukoku University, Japan from August 2007 to September 2010. He was a research scientist at Wave Engineering Laboratories, ATR Institute International, Japan from October 2010 to March 2014. He was an Assistant Professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan from April 2014 to March 2017. He became an assistant professor at Okayama University, Japan from April 2017. He is a guest research scientist at Wave Engineering Laboratories, ATR Institute International, Japan from October 2016. His research interest are communication systems, wireless networks, and signal processing. He received IEICE (the Institute of Electronics, Information and Communication Engineers) Communications Society Best Paper Award in 2016, 2020, and Best Tutorial Paper Award in 2017. Dr. Hou is a senior member of IEEE and IEICE.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

A Federated Cloud-Based Auction Mechanism for Real-Time Scheduling of Vehicle Sensors in Vehicle-Road-Cloud Collaborative System

Xueke DONG[†a)], **Wen TIAN**[†b)], **Xuyuan YE**[†], **Yining XU**[†], **Tiancheng WU**[†], ***and*** **Zhihao WANG**[†], ***Nonmembers***

SUMMARY Federated cloud, as a promising technology, can improve the computing capacity for autonomous driving in the vehicle-road-cloud collaborative system. However, the allocation of federated clouds should consider the environmental changes based on the real-time impact of vehicle terminal location. To improve computational efficiency while ensuring the effectiveness of federated clouds, this paper proposes a one-sided matching reverse auction based on the federated clouds (OSFC) method for scheduling autonomous driving sensors in a vehicle-road-cloud collaborative environment. This method dynamically allocates communication resources according to the actual situation of the vehicle terminals in real time. Numerical simulations show that our proposed OSFC method significantly improves computational efficiency while ensuring the effectiveness of federated clouds compared with state-of-the-art work.

key words: *automatic driving, federated clouds, onboard sensors, RSU, reverse auction, one-sided matching*

1. Introduction

Recently, traffic safety has become a growing concern for people worldwide. The association for safe international road travel states that road crashes account for 2.2% of all deaths globally. It is predicted that road injuries will become the fifth leading cause of death by 2030 [1]. The development of the Internet of Vehicles (IoV) helps improve traffic safety, efficiency, and driving experience. Through vehicle-to-vehicle communication, real-time traffic information can be shared, providing navigation suggestions and traffic congestion alerts to assist drivers in selecting the best routes. As an important technology in the IoV, the emergence of autonomous driving has become a hope for addressing personnel shortages and traffic safety issues. Autonomous driving vehicles at Level 4 can operate with only one driver in the passenger seat for supervision, while Level 5 autonomous driving vehicles can operate without a driver at all, significantly reducing the demand for personnel and the number of casualties in traffic accidents.

Autonomous driving can be broken into three components: algorithms, client systems, and the cloud platform [2]. Algorithms include sensing, perception, and decision. Autonomous driving vehicles utilize sensing systems such as radar and sonar to perceive the surrounding environment and make decisions based on the sensed information. An autonomous driving vehicle is a mobile terminal, while a standalone onboard system cannot adapt to different environments. The cloud platform can help store and process large amounts of data to assist the vehicle terminals in merging the perception information and providing control commands. During the journey of an autonomous driving vehicle, the onboard sensors perceive the surrounding environment, and wireless communication with the cloud is established through the roadside units (RSUs). The cloud analyzes and processes the data to make more accurate and reliable decisions for the autonomous driving vehicle. However, with the gradual expansion of the scope of the IoV, a single cloud is no longer able to meet the increasing computational requirements of autonomous driving. Therefore, the concept of federated clouds has emerged. The federated clouds combine multiple clouds to form a unified and collaborative cloud computing platform, enabling resource sharing and meeting the demands of autonomous driving. However, the resource allocation in the federated clouds is not fixed and needs to consider the real-time impact of the surrounding environment where the vehicle terminal is located. For example, when the vehicle is driving on a wide road with fewer surrounding vehicles, the clouds only need to access a portion of the onboard terminal's sensors, resulting in fewer communication resources required. On the other hand, when the vehicle is driving on a narrow road with more surrounding vehicles, the clouds need to access the majority of the vehicle terminal's sensors, leading to an increase in the required communication resources. As autonomous vehicles travel around a city, each second over 2GB of raw sensor data can be generated [2]. Therefore, if the clouds cannot use a dynamic adjustment mechanism, a large number of sensors in the vehicle terminals will work in real-time, resulting in resource waste and increased operating costs for the cloud. To improve the utilization of communication resources, it is necessary to dynamically adjust the resource allocation strategy based on the actual situation of the vehicle terminals in real time.

To the best of our knowledge, although there are many studies currently on the allocation of resources in vehicle cloud computing [3]–[5], most of them focus on the allocation of resources in an individual cloud, with few in the

Manuscript received January 10, 2024.
Manuscript revised April 24, 2024.
Manuscript publicized June 28, 2024.

[†]Key Laboratory of Intelligent Support Technology In Complex Environment, Nanjing University of Information Science and Technology, Ministry of Education, China.
a) E-mail: xkdong2002@163.com
b) E-mail: csusttianwen@163.com (Corresponding author)

DOI: 10.23919/transcom.2024EBP3010

context of federated clouds [8]–[10]. Especially, in these few articles related to federated clouds, the authors focus only on the bidirectional matching problem between the federated clouds and the vehicle terminals. However, in practical scenarios, there exists a three-way matching involving the federated clouds, vehicle terminals, especially the onboard sensors, and the RSUs.

Roadside unit (RSU) is an important component of the vehicle-road-cloud cooperative perception system. During the communication between the vehicle terminal and the clouds, RSU can act as a wireless relay communication device, significantly reducing the communication delay between the vehicle terminal and the clouds. Similar to the transmission of sensor data from the vehicle terminal, RSU also needs to be dynamically allocated communication resources based on the size and type of transmitted data. Therefore, RSUs should also be regarded as a part of the resource allocation problem. As a result, it is necessary to find a solution to solve the allocation problem among the federated clouds, onboard sensors, and RSUs.

Recently, auction theory has been widely used as a resource optimization allocation method in the field of cloud computing. Therefore, auction algorithms have become an effective method to dynamically call onboard sensors in the context of federated cloud collaboration. In this paper, we propose a one-sided matching algorithm (OSM) in our OSFC method, and the participants in the auction include the federated clouds, onboard sensors, and the RSUs.

This paper investigates the resource allocation problem of onboard sensors for autonomous vehicles in the federated cloud environment. In this scenario, the federated clouds dynamically utilize the sensors of vehicle terminals and RSUs based on the specific environment, thereby improving resource utilization. To the best of our knowledge, this paper is one of the first to propose an efficient resource allocation approach using auction algorithms among the federated clouds, onboard sensors, and RSUs. The main contributions of this paper can be summarized as follows:

- This paper proposes a OSFC method in the vehicular networking environment, utilizing RSUs as wireless relay devices and dynamically scheduling onboard sensors with a reverse auction model. The model involves three different participants: the federated clouds, onboard sensors, and RSUs. The federated clouds are responsible for acquiring sensor data from onboard sensors and allocating communication resources, while onboard sensors provide sensing data, and RSUs provide wireless relay communication services. The auction model is formalized as a 0-1 integer programming problem with the objective of maximizing the revenue of the federated clouds.
- Our auction aims to maximize the revenue for the federated clouds and takes a more favorable stance towards the federated clouds. Therefore, in our OSFC method, we propose a one-sided matching auction mechanism, which allows the federated clouds and the third-party auctioneer to unilaterally select the onboard sensors and RSUs. Through this one-sided matching auction mechanism, we can obtain a suboptimal auction solution with high computational efficiency.
- We conduct a testbed with varying numbers of federated clouds, onboard sensors, and RSUs to verify our proposed OSFC. The experimental results show that our method can improve computational efficiency while ensuring the effectiveness of federated clouds compared with state-of-art work, including VCG, FOGA, and RSBM.

The rest of this paper is structured as follows. Section 2 discusses the related work of the study. Section 3 provides system model, and Sect. 4 gives the details of the proposed method. Numerical simulations are conducted in Sect. 5. Finally, the conclusion of the article is given in Sect. 6.

2. Related Work

Although there have been many existing efforts dedicated to resource allocation in an individual cloud, such as [3]–[5], only a few have focused on resource allocation in the federated clouds environment [6], [7], especially in the context of vehicle-cloud collaboration [8]–[10]. In [8], the authors investigate the dynamic resource allocation problem for hosting latency-sensitive vehicle services in a federated cloud, aiming to maximize the number of served requests by meeting their delay requirements while minimizing VM migrations. However, this work does not consider the real-time changes in the environment during vehicle movement and the real-time interaction between vehicle terminals and cloud platforms. In [9], the authors formulate the problem of dynamic resource allocation for vehicular applications in the federated cloud as an optimization problem with the objective of minimizing the cost of the service provider, while meeting the delay requirements of the applications. However, this article does not consider the impact of the data volume transmitted between autonomous vehicles and the federated clouds on costs. In [10], the authors propose and implement a resource-based clustering algorithm scheme that groups vehicles with similar mobility and resource characteristics to form a dynamic federated vehicular cloud. While the role of RSUs in the IoV is considered in this article, the cost impact of RSUs as wireless relay communication devices is ignored.

In recent years, auction algorithms have been widely applied as an effective resource allocation method in cloud environments [11]–[13], [19]. In [11], the author proposes the PCAD auction method, which effectively addresses the resource allocation problem across data centers and reduces energy consumption in cloud computing data centers. In [12], the author proposes the ADAM data management method based on auction theory to better manage the demand and supply of cloud services in relation to vehicle parking issues. [13] presents a multi-unit double auction mechanism that effectively selects cloud alliances for users. While these auc-

tion mechanisms consider the sharing of resources in clouds, their purpose is to benefit the buyers of cloud services, which differs from the objective of this article. [19] proposes a Dynamic Combinatorial Double Auction (DCDA) model to improve social welfare and resource utilization. Although this work considers the varying resource demands of each cloud user, the method proposed in the article is only applicable to the auction between cloud users and cloud resource providers, and is not suitable for the three-party auction proposed in our paper.

RSU, as a wireless relay communication device, is an indispensable component in the vehicle-road-cloud cooperative perception system. [14] proposes an RSU-assisted VANETS vehicle resource search and cloud construction mechanism, where the involvement of RSU helps to avoid the limited search range caused by single-hop search and the frequent interruptions caused by multi-hop search. In [10], RSUs are considered as roadside clouds to initiate clustering and select the cluster head in the infrastructure-based clustering model. [17] proposes roadside unit (RSU)-assisted hybrid emergency message broadcasting (RA-HEMB) protocol for two-way grid roads in urban VANETs to improve the efficiency and reliability of the emergency message broadcasting. However, RSUs are merely used as auxiliary tools in these works, and none of these works have considered the three-way matching involving the federated clouds, the onboard sensors, and the RSUs in the vehicle-road-cloud cooperative perception system. Therefore, this paper proposes the OSFC method to solve the three-way matching problem.

3. System Model

As shown in Fig. 1, during the driving process, autonomous vehicles utilize sensors such as radar and navigation to gather information about the road conditions. They upload specific environmental information to the cloud platform through RSUs which act as relay communication devices. The vehicle terminal needs to adapt to the road conditions and gather environmental information by utilizing different sensors for different road condition information, such as vehicle density, spacing, and ambient temperature. The federated clouds then analyze the road condition information uploaded by the vehicle terminal and awaken different onboard sensors according to their needs. Specifically, in order to ensure real-time performance, our model considers that a federated cloud only interacts with one RSU at a single moment, but a cloud platform can choose another RSU at another moment [18].

Considering an autonomous vehicle driving on the road, different onboard sensors have different costs and performance, resulting in varying communication bandwidth usage. Similarly, as relay communication devices, RSUs also consume different sizes of communication resources due to differences in performance. Therefore, in the matching algorithm, each onboard sensor and each RSU need to provide their different demands for communication resources to facilitate resource scheduling by the federated clouds. Since both the onboard sensors and RSUs will ultimately receive communication resource rewards allocated by the cloud platforms, we can consider the resource scheduling process of the cloud platforms as a three-party reverse auction. Figure 2 illustrates the auction model of this approach, where the onboard sensors, as one party submitting communication demands, provide sensor data, the RSUs, as another party submitting communication demands, provide relay communication services, and the federated clouds, as the resource allocation party, provide communication resource rewards.

Since the communication resources required by the onboard sensors and RSUs are associated with their costs, we convert the occupied communication resources into consumed costs for ease of description. We consider the onboard sensors' demand for communication resources when transmitting data as the bid price of the onboard sensors, and the RSUs' demand for communication resources when forwarding data as the bid price of the RSUs. The federated clouds evaluate the required communication resource cost for data transmission based on their demand for communication data, which serves as the federated clouds' valuation. In addition, to ensure the efficiency of the auction, we introduce a virtual auctioneer to manage the auction process.

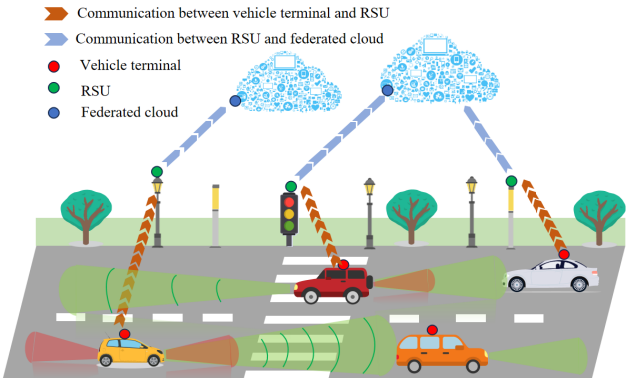


Fig. 1 Vehicle-road-cloud collaborative model.

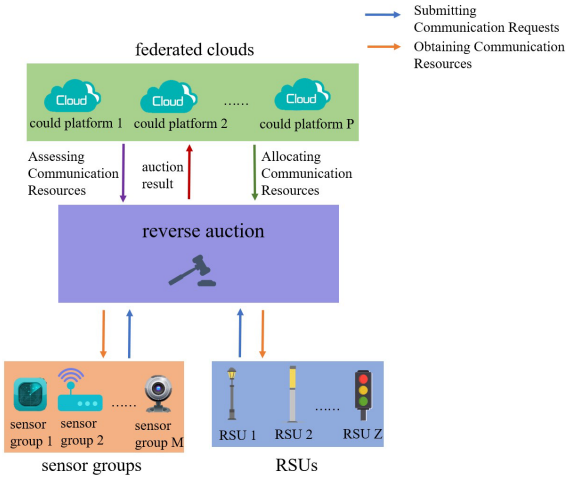


Fig. 2 Auction model.

4. Details of the Proposed Auction

This section introduces the details of the proposed auction, including the auction participants, the problem formulation and the one sided matching auction algorithm.

4.1 Auction Participants

4.1.1 Communication Resource Cost of Onboard Sensors

For onboard sensors, we divide them into multiple sensor clusters based on certain requirements (such as different manufacturers). Each sensor cluster corresponds to a group of onboard sensors, and each cluster contains different types of data information required by the federated clouds. The cost of onboard terminals in acquiring and transmitting data includes two aspects: the cost of acquiring different data s^d and the cost of transmitting different data s^{tx}. The relationship between the bandwidth occupied during data transmission and the transmission rate is summarized by the Shannon formula (1),

$$B = \frac{R}{\log(1 + S/N)}, \tag{1}$$

where B is the bandwidth, R is the maximum information transmission rate, and S/N is the signal-to-noise ratio. Assuming that there are M groups of onboard sensors, $\mathcal{M} = \{1, \ldots, m, \ldots, M\}$, each group of sensors can provide P data sets required by P cloud platforms, $\mathcal{P} = \{1, \ldots, p, \ldots, P\}$. For the sensor group m, the size of P data sets provided by it is represented by $\vec{d}_m = \left(d_{m,1}, \ldots\ldots, d_{m,p}, \ldots\ldots d_{m,P}\right)$, the transmission time is represented by $\vec{t}_m = \left(t_{m,1}, \ldots\ldots, t_{m,p}, \ldots\ldots, t_{m,P}\right)$, and the rate at which data is transmitted is represented by $\vec{R}_m = \left(R_{m,1}, \ldots\ldots, R_{m,p}, \ldots\ldots, R_{m,P}\right)$. Therefore, any relationship between $R_{m,p} \in \vec{R}_m$, $d_{m,p} \in \vec{d}_m$ and $t_{m,p} \in \vec{t}_m$ can be calculated by the following (2),

$$R_{m,p} = \Omega(d_{m,p}, t_{m,p}), \tag{2}$$

where $\Omega(\cdot)$ represents a function of two variables.

We use the following (3) to measure the cost of the communication resources of onboard sensors,

$$cost = \alpha B, \tag{3}$$

where $cost$ is the total cost of expenses, α is the unit bandwidth cost, and B is the occupied bandwidth.

For the same vehicle terminal, we assume that the transmission power when it sends data is constant S. As a result, the bandwidth $B_{m,p}$ required for data transmission can be calculated by (1) and (2), where $m \in \mathcal{M}$, $p \in \mathcal{P}$. Then submitting the bandwidth $B_{m,p}$ into Eq. (1) allows us to calculate the cost of transmitting data as (4),

$$s^{tx}_{m,p} = \alpha B_{m,p}. \tag{4}$$

Table 1 Notation table.

Notation	Defination
$P/Z/M$	Number of federated clouds/ RSUs/ sensor groups
$\mathcal{P}/\mathcal{Z}/\mathcal{M}$	Set of federated clouds/ RSUs/ sensor groups
$\vec{d}_m$	Date size vector of sensor group m
$\vec{t}_m$	Transmission time vector of sensor group m
$\vec{R}_m$	Transmission rate vector of sensor group m
$B_{m,p}$	Bandwidth required by sensor group m when transmitting data to cloud p
$s^{tx}_{m,p}$	Transmission cost of sensor group m for cloud p
$s^{d}_{m,p}$	Data cost of sensor group m for cloud p
$\sigma_{m,p}$	Unit data cost of sensor group m for cloud p
$s_{m,p}$	Total cost of sensor group m for cloud p
$\vec{s_m}$	Total cost vector of sensor group m
$\vec{q_m}$	Bid vector of sensor group m
$r^{tx}_{z,(m,p)}$	Transmission cost of RSU z
λ_z	Unit transmission distance cost of RSU z
$g_{z,m}$	Distance between the sensor group m and the RSU z
$r^{td}_{z,(m,p)}$	Delay cost of RSU z
ξ_z	Unit delay cost of RSU z
$d_{z,(m,p)}$	Size of data forwarded by RSU z
u_z	Data forwarding speed of RSU z
$r_{z,(m,p)}$	Total cost of data forwarding by RSU z
R_z	Total cost matrix of RSU z
H_z	Bid matrix of RSU z
$J_{p,(m,z)}$	Joint bid of sensor-RSU pair for cloud p
v_p	Valuation of cloud p
d_p	Data size required by cloud p
$x_{p,m,z}$	Match result of (p, m, z)
X	Match result matrix
$D^{w}_{(m,z)}$	Total payment for the winning sensor-RSU pair (m, z)
D^{s}_{m}	Payment for the winning onboard sensor group m
D^{r}_{z}	Payment for the winning RSU z
B^{win}_{m}	Winning bandwidth for sensor group m
L_A	Preference list of auctioner
L_p	Preference list of cloud p

The cost of acquiring data by a sensor is dependent on the performance of the sensor itself and the size of the collected data. As mentioned above, we divide all the sensors of the onboard terminal into M sensor groups. For each sensor group m, the cost incurred in collecting data p can be represented by the following (5),

$$s^{d}_{m,p} = \sigma_{m,p} d_{m,p}, \tag{5}$$

where $\sigma_{m,p}$ represents the unit cost of collecting data p by sensor group m, and $d_{m,p}$ represents the size of data p collected by sensor group m. Therefore, the communication resource cost of onboard sensors can be represented by the following (6),

$$s_{m,p} = s^{tx}_{m,p} + s^{d}_{m,p}. \tag{6}$$

The communication cost price of sensor group m can be represented by vector $\vec{s_m} = (s_{m,1}, s_{m,2} \ldots s_{m,p} \ldots s_{m,P})$. In addition, we use $\vec{q_m} = (q_{m,1}, q_{m,2} \ldots q_{m,p} \ldots q_{m,P})$ to represent the bid of sensor group m.

4.1.2 Communication Resource Cost of the RSUs

The RSUs provide relay services, which can be divided into transmission cost r^{tx} and delay cost r^{td}. Considering a set of RSUs $\mathcal{Z} = \{1, 2 \ldots, z, \ldots, Z\}$, the transmission cost of RSU z is r^{tx}_z, and the delay cost of it is r^{td}_z. The transmission cost is related to the distance $g_{z,m}$ between the vehicle terminal

and the RSU, which can be represented by the following,

$$r^{tx}_{z,(m,p)} = \lambda_z g_{z,m}, \tag{7}$$

where λ_z represents the unit transmission distance cost when the vehicular terminal communicates with the RSU z.

The delay cost r^{td} is related to the size of data forwarded by RSU, and it can be evaluated by the following,

$$r^{td}_{z,(m,p)} = \xi_z \frac{d_{z,(m,p)}}{u_z}, \tag{8}$$

where ξ_z represents the unit delay cost of RSU z, $d_{z,(m,p)}$ represents the size of data forwarded by RSU z, and u_z represents the data forwarding speed of RSU z.

Accordingly, the total cost of data forwarding by RSU z is given by,

$$r_{z,(m,p)} = r^{tx}_{z,(m,p)} + r^{td}_{z,(m,p)}. \tag{9}$$

Therefore, the total cost of RSU can be represented as a matrix $r_{z,(m,p)} \in R_z$. When RSUs participate in an auction, they submit a bid price matrix H_z of size $Z * M * P$, where $h_{z,(m,p)} \in H_z$ represents the bid of RSU z when it is matched with the onboard sensor group m to transmit data to the cloud platform p.

To facilitate analysis, let's assume that a cloud platform only selects one RSU and obtains the required communication data from one onboard sensor group. Unlike the traditional bilateral auction, this auction involves three parties, so we consider a pair consisting of an onboard sensor group and a RSU, and treat their total bid as a joint bid in order to reduce complexity.

Definition 1: (Joint bid) For the onboard sensor group m and RSU z, they will be referred to as the sensor-RSU pair (m, z), and the combined bid for providing data and services to the cloud platform p will be referred to as the joint bid $J_{p,(m,z)}$, where $J_{p,(m,z)} = q_{m,p} + h_{z,(m,p)}$. Therefore,the problem from the original three-party auction is changed into a two-way auction.

4.1.3 Valuation of Federated Clouds

This method considers the scenario of multiple cloud platforms, where each cloud platform processes different types of data.

Each cloud platform analyzes the environment, provides feedback to the auctioneer regarding data requirements, and estimates the value of the data based on these requirements. Considering there are P cloud platforms $\mathcal{P} = \{1, 2 \ldots, p, \ldots, P\}$. Each cloud platform needs to utilize the RSU's forwarding service to obtain sensor data to meet its computational requirements. The valuation is dependent on the data size d_p and data quality η_p required by each cloud platform. The formula for the estimation is given by the following,

$$v_p = \Psi(d_p, \eta_p), \tag{10}$$

where $\Psi(\cdot)$ represents a monotonic increasing function.

Data quality is closely related to accuracy, and can be described by the following (11) according to the relationship between testing accuracy and data size presented in [15],

$$v = \beta_1 * (1 + \beta_2 * d), \tag{11}$$

where β_1 and β_2 represent the parameters of the positive curve fitting.

Assuming that the federated clouds require P types of data, meaning there are a total of P cloud platforms, and the data size required by cloud platform p is given as d_p, with the corresponding estimated cost as v_p. Therefore, the estimated cost of all cloud platforms can be represented by $V = \{v_1, v_2, \ldots, v_p, \ldots, v_P\}$.

4.2 Problem Formulation

We use binary variables $x_{p,m,z}$ to represent whether the cloud platform successfully matches with the onboard sensor group and RSU. $x_{p,m,z} = 1$ represents a successful match, while $x_{p,m,z} = 0$ represents an unsuccessful match. The matrix $X = \{x_{p,m,z} | p \in \mathcal{P}, m \in \mathcal{M}, z \in \mathcal{Z}\}$ is used to represent the final matching result. The overall profit that the cloud platform can obtain can be expressed by the following,

$$E(X) = \sum_{p \in P} \sum_{m \in M} \sum_{z \in Z} x_{p,m,z}(v_p - q_{m,p} - h_{z,m,p}). \tag{12}$$

Therefore, the problem is transformed into finding the maximum value of the above equation,

$$\max_X E(X), \tag{13}$$

which is subject to the following constraint,

$$s.t. \quad x_{p,m,z} \in \{0, 1\}, \forall x_{p,m,z} \in X, \tag{13.1}$$

$$\sum_{m \in \mathcal{M}} x_{p,m,z} \leq 1, \forall p \in \mathcal{P}, \tag{13.2}$$

$$\sum_{p \in \mathcal{P}} x_{p,m,z} \leq 1, \forall m \in \mathcal{M}, \tag{13.3}$$

$$\sum_{p \in \mathcal{P}} \sum_{m \in \mathcal{M}} x_{p,m,z} \leq 1, \forall z \in \mathcal{Z}, \tag{13.4}$$

$$d_{m,p} \geq d_{z,m,p} \geq d_p, \forall p \in \mathcal{P}, \forall m \in \mathcal{M}, \forall z \in \mathcal{Z}. \tag{13.5}$$

(13.1) ensures that binary variables can only take values of 0 and 1, while (13.2) and (13.3) guarantees a one-to-one matching between the onboard sensor group and the cloud platform. This means that a cloud platform can only obtain the required data from one group of onboard sensors, and one group of onboard sensors can only provide data to one cloud platform. (13.4) ensures that only one RSU is used to serve one group of onboard sensors and one cloud platform. (13.5) ensures that the data provided by the onboard sensor group is not less than the data transmitted by the RSU, and the data transmitted to the cloud platform is also not less than the required communication data size of the cloud platform.

4.3 One Sided Matching Reverse Auction

In this section, a computationally efficient one sided matching auction algorithm is introduced by considering the relationships between federated clouds, onboard sensors, and RSUs.

4.3.1 Mechanism Design

In reverse auctions, to maximize the profit of the federated clouds, the auctioneer and the federated clouds have more initiative. This method, which considers the matching problem from a one sided perspective, is referred to as the one sided matching auction algorithm. The auctioneer tends to select combinations of cloud platform, sensor group, and RSU with higher profits, while the cloud platform tends to choose pairs of sensor group and RSU with lower joint bids. This leads to the concept of preference value, which is defined for combinations by the following,

$$Q_{p,(m,z)} = v_p - J_{p,(m,z)}. \tag{14}$$

After receiving the bids of the onboard sensor groups and RSUs, as well as the values of the federated clouds, the third-party auctioneer calculates the preference values for each combination of cloud platform, sensor group, and RSU. These values are then sorted in a non-ascending order. For example, if $Q_{p,(m,z)} > Q_{p',(m',z')}$, group $p,(m,z)$ will be placed ahead of $p',(m',z')$. Following this rule, a final list L_A will be obtained. Each cloud platform will also generate their respective lists in non-decreasing order based on the preference values when matching with different sensor-RSU pairs. For example, for cloud platform p, if $Q_{p,(m,z)} > Q_{p,(m',z')}$, sensor-RSU pair (m,z) will be placed ahead of pair (m',z') in the list L_p of the cloud platform p. In the end, there will be P individual preference lists for the federated clouds.

To prevent negative returns, we remove items with preference values less than zero in L_A and L_p. In addition, we add an empty sensor-RSU pair at the bottom of each cloud platform's preference list to ensure authenticity. This means that in case of auction failure, the cloud platform will be matched with an empty sensor-RSU pair.

According to the auctioneer and cloud platform's preference lists, we match the cloud platforms with suitable sensor-RSU pairs. It is important to note that not all cloud platforms will be matched with the sensor-RSU pairs with the highest preference value. For example, when a sensor-RSU pair is ranked first in both the list L_p of the cloud platform p and the the list L'_p of the cloud platform p', the sensor-RSU pair should be allocated to the appropriate cloud platform according to the inequality relationship between $Q_{p,(m,z)}$ and $Q_{p',(m,z)}$ in the auction's list L_A. In this case, the cloud platform with a lower preference value in L_A can only be matched with the next pair in its own list. Therefore, we match sensor-RSU pairs to each cloud platform following the above principle until all cloud platforms are matched with suitable sensor-RSU pairs.

Furthermore, the authenticity and individual rationality possessed by this method have been verified in [16], thus the bidding price can be regarded as equal to the cost price.

4.3.2 Distribution Rule

Definition 2: (Suboptimal Preference Value)We refer to the sensor-RSU pair in the cloud platform preference value list that ranks second only to the winning sensor-RSU pair as the suboptimal sensor-RSU pair. The preference value associated with it is referred to as the suboptimal preference value. For the cloud platform, the definition of the suboptimal preference value is as follows,

$$Q^{sub}_{p,(m,z)^{sub}} = v_p - J_{p,(m,z)^{sub}}, \tag{15}$$

where $(m,z)^{sub}$ represents the suboptimal sensor-RSU pair in the preference value list L_p of the cloud platform p, ranking second only to the winning sensor-RSU pair.

When the cloud platform p successfully matches with the sensor-RSU pair (m,z), the total payment received by the winning sensor-RSU pair is defined by the following,

$$D^w_{(m,z)} = v_p - Q^{sub}_{p,(m,z)^{sub}}. \tag{16}$$

Accordingly, the payment received by the winning onboard sensor group is given by the following,

$$D^s_m = (v_p - Q^{sub}_{p,(m,z)^{sub}})\frac{q_{m,p}}{J_{p,(m,z)}}, \tag{17}$$

and the payment received by the winning RSU is given by,

$$D^r_z = (v_p - Q^{sub}_{p,(m,z)^{sub}})\frac{h_{z,m,p}}{J_{p,(m,z)}}. \tag{18}$$

4.3.3 Allocation of Communication Resources

After the auction is completed, the matched sensor groups are the ones that the federated clouds will invoke. Combining Eqs. (4), (5), and (6), the allocated bandwidth for transmitting data from the called sensor groups should be equivalent to the remaining payment after deducting the cost of compensating for the data acquisition, which is expressed as,

$$B^{win}_m = \frac{D^s_m - s^d_{m,p}}{\alpha}. \tag{19}$$

The OSM algorithm is summarized in Algorithm 1. In Algorithm 1, the preference value for each combination of cloud platform, sensor group, and RSU is calculated in line 1 according to the input variables, the time complexity of which is $O(PMZ)$. Line 2–6, which sort L_A and L_p, is $O(PMZlog(PMZ))$ and $O(MZlog(MZ))$ [20]. Line 7–12 obtain the matching result and suboptimal sensor-RSU pair based on the preference list L_A and L_p. Line 13–19 allocate

Algorithm 1: one sided matching reverse auction algorithm(OSM).

Input: $\mathcal{P}$, $\mathcal{M}$, $\mathcal{Z}$, value of federated clouds $V = \{v_1, ..., v_p, ..., v_P\}$, bid of RSUs H_Z of size $P * M * Z$,bid of sensor groups $\{\vec{q_1}, ...\vec{q_m}, ...\vec{q_M}\}$

Output: Results of the matching , Allocation of communication resources

1 *Initialize the mat X,calculation of the joint bid J and the preference of the auctioneer* L_A;
2 Sort L_A,"Non-ascending order";
3 **for** $p \leftarrow 1$ **to** P **do**
4 | *Calculate the* L_p;
5 | *Sort the* L_p *,"Non-ascending order"*;
6 **end**
7 **for** $p \leftarrow 1$ **to** P **do**
8 | *Find* $Q_{a,b,c} == L_A(1)$;
9 | $L_A \leftarrow L_A \setminus \{$ *all elements related to a,b,c*$\}$;
10 | $x_{a,b,c} \leftarrow 1$;
11 | *Find* $(b,c)^{sub}$ *in* L_p *according to Def. 2*;
12 **end**
13 **for** $x_{p,m,z} \in X, (p \in \mathcal{P}, m \in \mathcal{M}, z \in \mathcal{Z})$ **do**
14 | **if** $x_{p,m,z} == 1$ **then**
15 | | *Calculate the payment* D_m^s *for the sensor group m according to (17)* ;
16 | | *Calculate the payment* D_z^r *for the RSU z according to (18)* ;
17 | | *Allocate communication resources for the onboard sensors according to (19)*;
18 | **end**
19 **end**
20 **return** X, D^s, D^r

the payment to the onboard sensors and RSUs based on the proposed distribution rule. Similarly, the complexity of line 7–19 is $O(2PMZ)$. Compared with the tripartite matching method for optimal matching, the complexity of which is $O(2^{PMZ})$ [16], OSM has polynomial time complexity.

5. Simulation

5.1 Simulation Setting

The valuation provided by the federated clouds is mainly related to the data required by them, so we use the variable $v_p = \beta_1 * (1 + \beta_2 * d_p)$ mentioned in (11) to represent the valuation of the federated clouds [15], where v_p represents the valuation of cloud platform p, and d_p represents the size of data the cloud platform p needs. We set $\beta_2 = 0.01$, and let the value of β_2 follow a uniform distribution $[4, 6]$, which is different for various cloud platforms. The size of the required data for the federated clouds follows a uniform distribution $[3000, 5000]$. For the onboard sensors, we suppose that data size is normalized by 500 units and that the normalized size data follows a uniform distribution $[12, 30]$. The unit cost of collecting data is randomly chosen from $[0.0002, 0.0006]$. Considering that the transmission time of P types of sensor data from the same sensor group to the cloud platform is the same, it is randomly selected from a uniform distribution $[0.02, 0.08]$ seconds [22]. Based on this, the transmission rate of sensor data can be evaluated. The cost per unit bandwidth is set to 0.02 and the signal-to-noise ratio for channel transmission is 20 dB [23]. As a result, the cost of onboard sensors can be calculated. For the RSUs, we assume that the unit transmission cost during interaction with the vehicle terminal follows a uniform distribution $[0.02, 0.06]$, and the transmission distance follows a distribution $[125, 250]$ meters [24]. For simplicity, we consider that the size of the sensor data sent by the onboard sensors is the same as the size of data received and forwarded by the RSU. Additionally, the forwarding rate of each RSU is randomly selected from the range $[40, 100]$, and the unit delay cost follows a uniform distribution $[0.5, 1]$ [25].

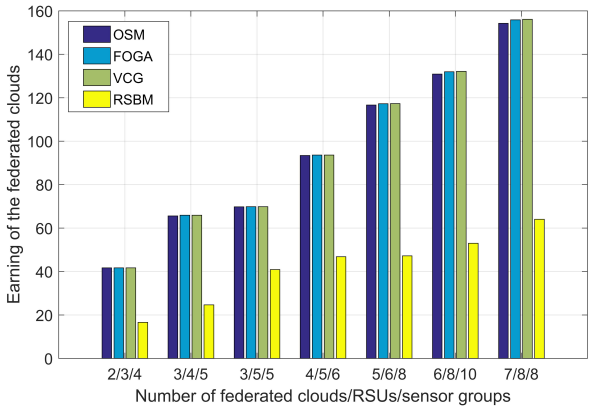


Fig. 3 The total earning of the federated clouds of different methods considering various federated clouds /RSUs/ sensor groups.

To validate the performance of the proposed algorithm in this paper, we compare our method (OSM) with existing baseline algorithms, which include fragmental optimization genetic algorithm (FOGA), Vickrey-Clarke-Groves (VCG) based optimal reverse auction, and random sampling-based method (RSBM). We conducted the simulation with MATLAB R2016a on Intel(R) Core(M) i5-13500E@2.6 GHz and the simulation results demonstrated in the following sections are the results averaged over 500 simulations.

5.2 Simulation Results

This section presents the simulation results and provides a brief description of the findings.

Figure 3 shows the total revenue of the federated clouds considering various federated clouds /RSUs/ sensor groups based on OSM and three other baseline algorithms. For example, when the number of federated clouds /RSUs/ sensor groups is 3/4/5, it means the problem is under 3 federated clouds, 4 RSUs and 5 sensor groups. It can be observed from Fig. 3 that, under the same number of federated clouds /RSUs/ sensor groups, our proposed OSM outperforms the RSBM algorithm by providing higher profits to the federated clouds. The performance of OSM, VCG and FOGA algorithms is basically the same when the number of federated clouds/ RSUs/ sensor groups is small. Although OSM slightly underperforms VCG and FOGA when the overall quantity is larger, our OSM demonstrates significantly higher

Table 2 Running time (seconds) of different methods considering various federated clouds /RSUs/ sensor groups.

Variable Value \ Methods	OSM	FOGA	VCG	RSBM
2/3/4	0.0030	0.1716	0.0238	0.0026
3/4/5	0.0441	0.5903	0.0533	0.0042
3/5/5	0.0040	2.3949	0.6121	0.0031
4/5/6	0.0112	6.1367	13.2762	0.0107
5/6/8	0.0429	15.5298	295.5817	0.0369
6/8/10	0.0459	88.2567	567109	0.0366
7/8/8	0.0460	2663.861	846972	0.0357

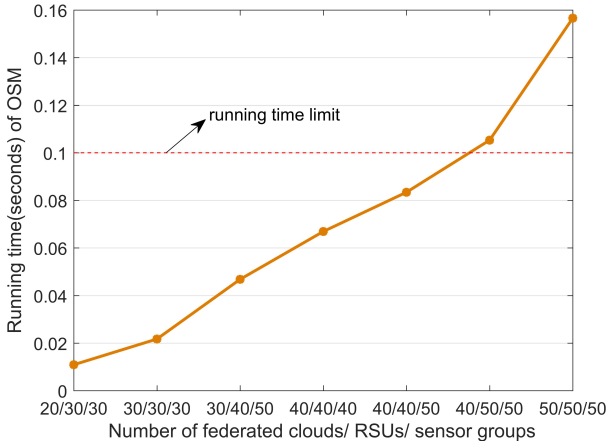


Fig. 4 Running time of OSM with various federated clouds /RSUs/ sensor groups.

computational efficiency compared to VCG and FOGA algorithms as shown in Table 2. When the number of federated clouds /RSUs/ sensor groups is 7/8/8, the running time of OSM is different from FOGA by 6 orders of magnitude and VCG by 8 orders of magnitude. As a result, our OSM can obtain a satisfying trade-off between total revenue and running time.

As shown in Fig. 4, the running time of OSM increases with the number of federated clouds /RSUs/ sensor groups increasing. Moreover, the running time of OSM exceeds 100 ms when the number of federated clouds /RSUs/ sensor groups is 40/50/50, while 3GPP NR-V2X supports a maximum latency of 100 ms [21]. As a result, the OSM proposed in this paper has a certain limit on the number of participants. In our simulation, the number of participants on each side should be less than 40–50 (the specific number limit depends on the hardware computing performance).

Figure 5 shows the bid and payment received by the winning RSUs and sensor groups under 10 federated clouds, 12 RSUs and 14 sensor groups. Joint bid and total payment of the sensor-RSU pairs, as well as the individual bid and payment of RSU or sensor group are all shown in Fig. 5. The horizontal coordinate is the identifier of the winning federated cloud /RSU/ sensor group. For example, 2/3/1 represents the cloud platform 2 is successfully matched with RSU 3 and sensor group 1. It can be observed that the payment received by RSUs and sensor groups are always greater than the total bid. In other words, both the sensor groups and

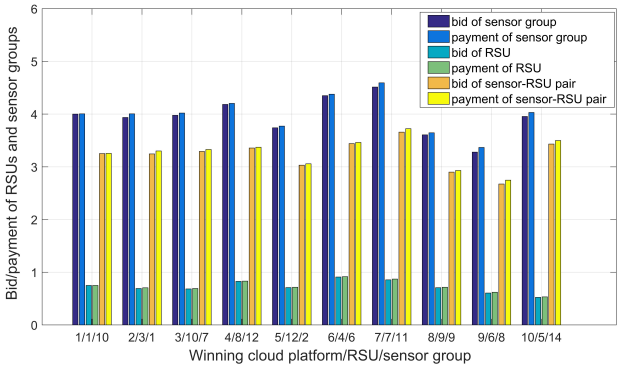


Fig. 5 The bid/payment of each group of winning cloud platform/RSU/sensor group under 10 federated clouds, 12 RSUs and 14 sensor groups.

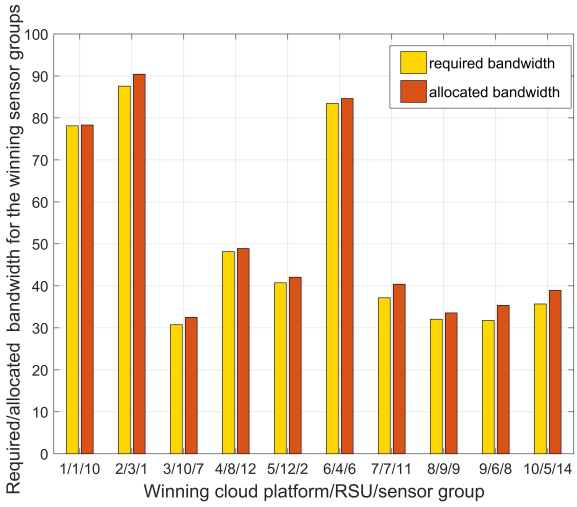


Fig. 6 The required/ allocated bandwidth for the winning sensor groups under 10 federated clouds, 12 RSUs and 14 sensor groups.

RSUs can obtain non-negative returns in the auction. This verifies that the sensor groups and RSUs have individual rationality under the proposed OSM [26]. Figure 6 illustrates the required bandwidth for each winning sensor group before the auction and the allocated bandwidth after the auction under 10 federated clouds, 12 RSUs and 14 sensor groups. Similarly, it can be observed that the allocated bandwidth is always greater than the required bandwidth, indicating the rationality of the proposed OSM in this paper.

6. Conclusions

In this paper, we propose an OSFC method for scheduling autonomous driving sensors in a vehicle-road-cloud collaborative environment based on a reverse auction algorithm. The participants in this OSFC method include the federated clouds, onboard sensors, and RSUs. Taking into account the number of participants, we propose the concept of sensor-RSU pairs to simplify the auction. We introduce a one-sided matching reverse auction(OSM) algorithm based on suboptimal solutions in our proposed OSFC method, which can

obtain relatively optimal matching results. We finally use extensive simulations to demonstrate that the OSFC method is more computationally efficient while ensuring the effectiveness of federated clouds compared with the baseline algorithms and verify that the allocation of communication resources using this method is reasonable.

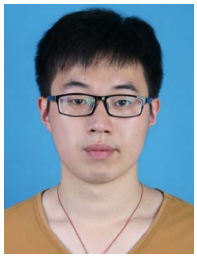
Acknowledgments

This work was supported by the National Key R&D Program of China (Grants No. 2021QY0700), the National Natural Science Foundation of China (Grants No. 62401269, U21B2003, 62072250), Jiangsu Province Natural Science Foundation (Grants No. BK20230415), NUIST Students' Platform for Innovation and Entrepreneurship Training Program (Grants No. XJDC202410300049), Jiangsu Province Students' Platform for Innovation and Entrepreneurship Training Program (Grants No. 202410300123Y) and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grants No. 23KJB120007).

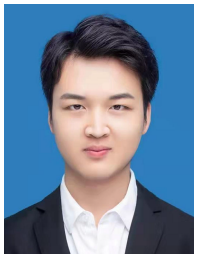
References

[1] Association for Safe International Road Travel, "Annual global road crash statistics," Online: http://www.asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics, Available on Oct. 14, 2018.
[2] S. Liu, L. Li, J. Tang, and S. Wu, Creating Autonomous Vehicle Systems, Morgan & Claypool Publishers, 2018.
[3] S. Yadav and M.M. Kaur, "Genetic algorithm-based data allocation in multi media using cloud computing," Proc. 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp.1668–1671, May 2023.
[4] K. Metwally, A. Jarray, and A. Karmouch, "MILP-based approach for efficient cloud IaaS resource allocation," Proc. IEEE 8th International Conference on Cloud Computing, pp.1058–1062, June-July 2015.
[5] J. Chen, "A cloud resource allocation method supporting sudden and urgent demands," Proc. Sixth International Conference on Advanced Cloud and Big Data (CBD), pp.66–70, Aug. 2018.
[6] J. Xu and B. Palanisamy, "Cost-aware resource management for federated clouds using resource sharing contracts," Proc. IEEE 10th International Conference on Cloud Computing (CLOUD), pp.238–245, June 2017.
[7] M.V. Haresh, S. Kalady, and V.K. Govindan, "Agent based dynamic resource allocation on federated clouds," Proc. IEEE Recent Advances in Intelligent Computational Systems, pp.111–114, Sept. 2011.
[8] M. Najm, M. Patra, and V. Tamarapalli, "An adaptive and dynamic allocation of delay-sensitive vehicular services in federated cloud," Proc. International Conference on COMmunication Systems & NETworkS (COMSNETS), pp.97–100, Jan. 2021.
[9] M. Najm, M. Patra, and V. Tamarapalli, "Cost-and-delay aware dynamic resource allocation in federated vehicular clouds," IEEE Trans. Veh. Technol., vol.70, no.6, pp.6159–6171, June 2021.
[10] W.M. Danquah and D.T. Altilar, "UniDRM: Unified data and resource management for federated vehicular cloud computing," IEEE Access, vol.9, pp.157052–157067, Nov. 2021.
[11] Y. Mao, X. Xu, L. Wang, and P. Ping, "Priority combinatorial double auction based resource allocation in the cloud," Proc. IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), pp.224–228, Aug. 2020.
[12] S.R. Rizvi, S. Zehra, S. Olariu, and S. El-Tawab, "ADAM: An auction-based datacenter management in vehicular cloud," Proc. IEEE International Conference on Smart Mobility (SM), pp.138–143, March 2023.
[13] A.I. Middya, B.K. Ray, and S. Roy, "Auction-based resource allocation mechanism in federated cloud environment: TARA," IEEE Trans. Services Comput., vol.15, no.1, pp.470–483, Jan.-Feb. 2022.
[14] Y. Lee, H. Choi, Y. Nam, S. Park, and E. Lee, "RSU-driven cloud construction and management mechanism in VANETs," Proc. IEEE 8th International Conference on Cloud Networking (CloudNet), pp.1–4, Nov. 2019.
[15] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo, "A learning-based incentive mechanism for federated learning," IEEE Internet Things J., vol.7, no.7, pp.6360–6368, July 2020.
[16] Z. Cheng, M. Liwang, X. Xia, M. Min, X. Wang, and X. Du, "Auction-promoted trading for multiple federated learning services in UAV-aided networks," IEEE Trans. Veh. Technol., vol.71, no.10, pp.10960–10974, Oct. 2022.
[17] D. Xi, H. Zhang, Y. Cao, and D. Yuan, "An RSU-assisted hybrid emergency message broadcasting protocol for VANETs," IEEE Internet Things J., vol.10, no.19, pp.17479–17489, Oct. 2023.
[18] H. Zhou, M. Li, P. Sun, B. Guo, and Z. Yu, "Accelerating federated learning via parameter selection and pre-synchronization in mobile edge-cloud networks," IEEE Trans. Mobile Comput., March 2024.
[19] Q. Li, X. Jia, C. Huang, and H. Bao, "A dynamic combinatorial double auction model for cloud resource allocation," IEEE Trans. Cloud Comput., vol.11, no.3, pp.2873–2884, July-Sept. 2023.
[20] Z. Zheng, Y. Gui, F. Wu, and G. Chen, "STAR: Strategy-proof double auctions for multi-cloud, multi-tenant bandwidth reservation," IEEE Trans. Comput., vol.64, no.7, pp.2071–2083, July 2015.
[21] 3GPP TR 22.886(V16.2.0), "Study on enhancement of 3GPP support for 5G V2X services," Dec. 2018.
[22] M. Najm, M. Patra, and V. Tamarapalli, "An adaptive and dynamic allocation of delay-sensitive vehicular services in federated cloud," Proc. International Conference on COMmunication Systems & NETworkS (COMSNETS), pp.97–100, Jan. 2021.
[23] W.G. Cassidy, N. Jaber, S.A. Ruppert, J. Toimoor, K.E. Tepe, and E. Abdel-Raheem, "Interference modelling and SNR threshold study for use in vehicular safety messaging simulation," Proc. 26th Biennial Symposium on Communications (QBSC), pp.52–55, May 2012.
[24] H. Gao, C. Liu, Y. Li, and X. Yang, "V2VR: Reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability," IEEE Trans. Intell. Transp. Syst., vol.22, no.6, pp.3533–3546, June 2021.
[25] P. Li, Q. Liu, C. Huang, J. Wang, and X. Jia, "Delay-bounded minimal cost placement of roadside units in vehicular ad hoc networks," Proc. IEEE International Conference on Communications (ICC), pp.6589–6594, June 2015.
[26] Z. Gao, M. Liwang, S. Hosseinalipour, H. Dai, and X. Wang, "A truthful auction for graph job allocation in vehicular cloud-assisted networks," IEEE Trans. Mobile Comput., vol.21, no.10, pp.3455–3469, Oct. 2022.

Xueke Dong graduated from Xishan Senior High School, Wuxi, China, in 2021. She is currently pursuing a bachelor's degree in the School of Electronic and Information Engineering at Nanjing University of Information Science and Technology, Nanjing, China. Her currently research interests include wireless communication and Internet of Things.



Wen Tian received the B.S. degree in physics from Changsha University of Science and Technology, Changsha, China, in 2014, the M.S. degree in control theory and control engineering from the Jiangsu University of Science and Technology, Zhenjiang, China, in 2017, and the Ph.D. degree in control science and engineering from Nanjing University of Science and Technology, Nanjing, China. He is currently a Lecturer with the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing. His research interests include cyber–physical systems, Internet of Things, game theory, and covert communication.



Xuyuan Ye graduated from Yangzi High School Affiliated to Nanjing Normal University, Nanjing, China, in 2021. He is currently pursuing a bachelor's degree in the School of Electronic and Information Engineering at Nanjing University of Information Science and Technology, Nanjing, China. His currently research interests include remote sensing and convolutional neural network.

Yining Xu is currently pursuing a bachelor's degree in the School of Electronic and Information Engineering at Nanjing University of Information Science and Technology, Nanjing, China. His currently research interests include covert communication and UAV communication.

Tiancheng Wu graduated from Xishan Senior High School, Wuxi, China, in 2022. He is currently pursuing a bachelor's degree in the School of Electronic and Information Engineering at Nanjing University of Information Science and Technology, Nanjing, China. His currently research interests include covert communication and UAV communication.

Zhihao Wang is a sophomore majoring in communication engineering at Nanjing University of Information Science and Technology. He has a strong research interest in secure communications and unmanned communications.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

Effect of Binary Labeling Schemes on PAS with BICM-ID System Performance over the AWGN and Optical Fiber Channels

Mamoru KOMATSU[†a)], ***Student Member and*** **Akira NAKA**[†b)], ***Member***

SUMMARY Bit-interleaved coded modulation with iterative decoding (BICM-ID) effectively provides a high spectral efficiency and coding gain for digital coherent systems over additive white Gaussian noise (AWGN) and optical fiber channels. We previously proposed combining probabilistic amplitude shaping (PAS) with BICM-ID to further improve the system performance. However, the BICM-ID performance depends on the binary labeling scheme used for the constellation points. In this study, we evaluated the effect of binary labeling schemes on the performance of the PAS with BICM-ID system. Numerical simulations showed that the PAS with BICM-ID system employing a suitable binary labeling scheme offers a significant coding gain over both the AWGN and optical fiber channels. The system is also robust against performance degradation caused by the optical Kerr effect in the optical fiber channel. We used an extrinsic information transfer (EXIT) chart to analyze the suitability of binary labeling schemes and the effect of bit interleavers. The results showed that a binary labeling scheme is suitable if the slope of the demodulator's EXIT curve is close to the slope of the decoder's EXIT curve. The EXIT chart analysis also showed that inserting bit interleavers mitigates the performance degradation during iterative decoding. In addition, we used bitwise mutual information to evaluate the SNR penalty due to shaping gap and coding gap, and coding gain offered by iterative decoding of BICM-ID.

key words: *probabilistic amplitude shaping (PAS), bit-interleaved coded modulation with iterative decoding (BICM-ID), extrinsic information transfer (EXIT) chart, binary labeling, optical fiber transmission*

1. Introduction

In optical fiber communication, digital coherent systems play a key role in meeting the demand from increased Internet traffic [1]. One of the benefits of digital coherent systems is the employment of coded modulation, which combines high-order modulation formats and forward error correction (FEC) codes for high spectral efficiency. For example, bit-interleaved coded modulation (BICM) is a relatively simple system that inserts a bit interleaver between the FEC and modulation, but it provides a high spectral efficiency and high coding gain over additive white Gaussian noise (AWGN) channels [2], [3].

BICM with iterative decoding (BICM-ID) improves upon the original BICM by incorporating the turbo principle [4], [5]. In practical systems, FEC codes have SNR gap between "SNR achieving its capacity" and "actual SNR achieving error-free". The SNR gap is called as coding gap. BICM-ID systems have a potential to reduce coding gap by improving the reliability of demodulation with a-priori information fed back from decoder to demodulator on the receiver side [5]. Note that under the assumption of using ideal FEC code which has no coding gap, BICM-ID provides no gain except for modulation formats not having Gray labeling, e.g., 8-QAM.

The improvement in performance by BICM-ID depends on the binary labeling of constellation points. Therefore, choosing a suitable binary labeling scheme to acquire a high coding gain is important. Several approaches have been proposed to design the binary labeling scheme, such as a heuristic approach [6] and optimization methods like the binary switching algorithm (BSA) [7], [8] and genetic algorithm [9].

The spectral efficiency of coded modulation can be further increased by constellation shaping, which employs a Gaussian distribution according to Shannon's theory to achieve a shaping gain. There are two types of constellation shaping: geometric shaping (GS) [10]–[13] and probabilistic shaping (PS) [14]–[17]. GS has a uniform probability of symbol occurrence with non-equispaced constellation points while PS has a non-uniform probability of symbol occurrence on conventional equispaced constellation points. Recently, Böcherer et al. [14] proposed a practical system incorporating PS called probabilistic amplitude shaping (PAS) that has received attention owing to its shaping gain and compatibility with conventional FEC codes.

Several combinations of BICM-ID and constellation shaping have been proposed to obtain coding and shaping gains [18]–[22]. For example, Khoo et al. [18] proposed combining BICM-ID with PS. They realized PS by dividing 16-ary quadrature amplitude modulation (QAM) constellation points into three subsets based on power level and frequently transmitting symbols from lower-power subsets using an additional binary code. Arafa et al. [19] combined BICM-ID with GS and employed two types of non-equispaced constellations: rectangular and circular. Naka [20] proposed combining BICM-ID and PAS and evaluated the system by employing two types of binary labeling schemes for 64-QAM. However, bit interleavers were not inserted in the proposed system because of the unique modulation process of PAS. We [21] proposed improving this system by inserting bit interleavers and used an extrinsic information transfer (EXIT) chart to evaluate the bit error rate (BER) performance with all possible binary labeling schemes of 64-QAM and the effect of the interleavers. How-

Manuscript received January 12, 2024.
Manuscript revised April 5, 2024.
Manuscript publicized July 18, 2024.
[†]Graduate School of Science and Engineering, Ibaraki University, Hitachi-shi, 316-8511 Japan.
a) E-mail: 23nd303r@vc.ibaraki.ac.jp
b) E-mail: akira.naka.dr@vc.ibaraki.ac.jp
DOI: 10.23919/transcom.2024EBP3011

ever, this EXIT chart analysis [21] was specific to the effect of interleavers. A general EXIT chart analysis is needed to clarify the relationship between labeling and system performance. Additionally, the previous studies on combining BICM-ID and constellation shaping [18]–[22] only considered AWGN channels. For application to optical fiber systems, their performance over optical fiber channels needs to be considered, as well as the degradation due to the optical Kerr effect.

Here, we extend our previous study [21], and the main contributions of this paper are summarized as follows: (i) we investigate the relationship between binary labeling and the BER performance for the PAS with BICM-ID by using EXIT chart for all possible binary labeling schemes on 64-QAM format; (ii) we evaluate the insertion of random bit interleavers to the PAS with BICM-ID system by using EXIT chart. Additionally, we explain the effect of these interleavers by an example of the exchanged LLR sequences; (iii) we evaluate the performance of the PAS with BICM-ID system over the optical fiber channel and observe the degradation due to the optical Kerr effect. In addition, we show consensus of the suitable binary labeling for PAS with BICM-ID system over the AWGN and optical fiber channels by evaluating the degradation due to the optical Kerr effect with respect to each binary bit.

The remainder of this paper is organized as follows. Section 2 reviews the PAS with BICM-ID system model in detail. Section 3 shows all possible binary labeling schemes of 8-ary pulse amplitude modulation (PAM) and 64-QAM for the PAS system. Section 4 presents numerical simulations performed to evaluate the BER performance over the AWGN and optical fiber channels. In addition, analyses based on the EXIT chart and bitwise mutual information (MI) are presented in Sect. 4. Section 5 concludes the paper.

2. System Model

Consider M-PAM and M^2-QAM, where the number of bits in an in-phase (I) or quadrature (Q) PAM symbol is denoted by $m = \log_2 M$. Figure 1 shows a diagram of the PAS with BICM-ID system [21], which combines the original PAS (i.e., code rate $R_c = (m-1)/m$ system in [14]) and BICM-ID. On the transmitter side, the k-bit information bit sequence is denoted by $\boldsymbol{u} \in \{0,1\}^k$. $\boldsymbol{u}$ is input into a distribution matcher (DM), where $\boldsymbol{u}$ is transformed into an n-length amplitude sequence $\boldsymbol{a} \in \{1,3,\ldots,M-1\}^n$. This sequence is chosen to satisfy the Gaussian distribution for improved receiver sensitivity [14]–[16]. Moreover, we use enumerative sphere shaping (ESS) [16], which offers high shaping gain with short length of amplitude sequence. For ESS, we define the input bit length as k_s and output amplitude length as n_s. After the DM process, the amplitude-to-bit-labeling process converts the amplitude sequence $\boldsymbol{a}$ into the $(m-1)n$-bit amplitude bit sequence $\boldsymbol{b}$. This conversion is based on $m-1$ bits, which indicates the amplitude information in the binary labeling of M-PAM [21]. The amplitude bit sequence $\boldsymbol{b}$ is interleaved by a random bit interleaver

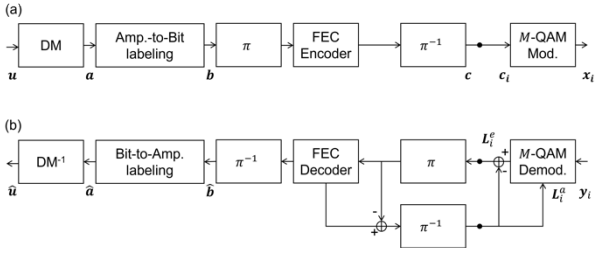


Fig. 1 PAS with BICM-ID system: (a) transmitter side and (b) receiver side.

(π) and is then fed into the FEC encoder. We previously proposed this interleaving and deinterleaving (π^{-1}) after the FEC encoder [21] to reduce BER degradation due to the non-uniform error correction of the pragmatic low-density parity-check (LDPC) decoder during the iterative decoding of BICM-ID. To retain the amplitude information formed in the DM at the modulator, the amplitude bit sequence $\boldsymbol{b}$ is interleaved before the FEC encoder and deinterleaved after the FEC encoder. Then, the coded and deinterleaved bit sequence $\boldsymbol{c}$ is divided into $2m$ bits that form the subsequence $\boldsymbol{c}_i = (c_i(1),\ldots,c_i(2m))$, where $i = 1,\ldots,n/2$. In the subsequences, $c_i(1)$ and $c_i(m+1)$ are parity bits indicating whether the binary labeling of M^2-QAM has a positive or negative sign, and the others are elements of $\boldsymbol{b}$ indicating the amplitude bits in the binary labeling of M^2-QAM. Each subsequence $\boldsymbol{c}_i$ is mapped to a QAM symbol $x_i \in X$, where X denotes a QAM symbol set.

On the receiver side, the coded and deinterleaved bit sequence $\boldsymbol{c}$ is estimated by the iterative decoding of BICM-ID. In the demodulator, the extrinsic LLR subsequence $\boldsymbol{L}_i^e = \left(L_i^e(1),\ldots,L_i^e(2m)\right)$ is calculated by using the received symbol y_i and corresponding a-priori LLR subsequence $\boldsymbol{L}_i^a = \left(L_i^a(1),\ldots,L_i^a(2m)\right)$ [8]:

$$L_i^e(j) = \log \frac{\displaystyle\sum_{x \in X_j^1} p(x)p\left(y_i \middle| x\right) p\left(x \middle| \boldsymbol{L}_i^a\right)}{\displaystyle\sum_{x \in X_j^0} p(x)p\left(y_i \middle| x\right) p\left(x \middle| \boldsymbol{L}_i^a\right)} - L_i^a(j) \quad (1)$$

where X_j^b is the subset of the QAM symbol set where the jth bit is $b \in \{0,1\}$, $p(x)$ is the probability mass function of x, and $p\left(y_i \middle| x\right)$ is a conditional probability density function y_i given one of the QAM symbol x. The conditional probability density function $p\left(y_i \middle| x\right)$ in AWGN channel is described by

$$p\left(y_i \middle| x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|y_i - x|^2}{2\pi\sigma^2}\right), \quad (2)$$

where σ^2 is the noise variance. In addition, we used the same conditional probability density function for simulation and analysis in the optical fiber channel (in Sects. 4.6–4.8). This assumes that the degradation due to the optical Kerr effect in the optical fiber channel is noise. It might be suboptimal for an optical fiber channel, but the assumption is reasonable because the optical fiber channel is not known in closed

form. $p\left(x\left|\boldsymbol{L}_i^a\right.\right)$ terms in (1) is calculated from a-priori LLR subsequence $\boldsymbol{L}_i^a$ and one of the QAM symbol x:

$$p\left(x\left|\boldsymbol{L}_i^a\right.\right) = \prod_{j=1}^{2m} \frac{\exp\left(\mu_j(x)L_i^a(j)\right)}{1+\exp\left(L_i^a(j)\right)}, \tag{3}$$

where $\mu_j(x)$ is a function that provides the j th bit in the QAM symbol x.

Note that a-priori LLR subsequence $\boldsymbol{L}_i^a$ is zero in the initial demodulation. The extrinsic LLR subsequences $\boldsymbol{L}_i^e$ are converted as corresponding bit sequence $\boldsymbol{c}$ before interleaving and feeding into the FEC decoder. Here, the iteration in the iterative decoding of the FEC decoder is denoted as the "inner iteration." The extrinsic LLR sequence of the FEC decoder that is fed back to the demodulator is regarded as a-priori LLR subsequence $\boldsymbol{L}_i^a$ for the iterative decoding of BICM-ID, which is denoted as the "outer iteration." After some outer iterations, the deinterleaved estimated amplitude bit sequence $\hat{\boldsymbol{b}}$ is converted to the estimated amplitude sequence $\hat{\boldsymbol{a}}$ by bit-to-amplitude labeling. Finally, the estimated information bit sequence $\hat{\boldsymbol{u}}$ is obtained from the estimated amplitude sequence $\hat{\boldsymbol{a}}$ by the inverse DM process.

3. Binary Labeling for the PAS System

We considered all possible binary labeling schemes of 8-PAM and 64-QAM for PAS systems. As mentioned in Sect. 2, PAS systems deal with 8-PAM symbols on I and Q elements independently on the 64-QAM constellation. Therefore, it is necessary to address the labeling schemes for 64-QAM as the combination of two labeling schemes for 8-PAM on I and Q elements independently. This implies that the binary label of 64-QAM for PAS cannot be jointly optimized on I and Q elements.

Since all labeling schemes for 64-QAM employing PAS are combinations of two labeling schemes of 8-PAM, we first needed to consider all possible binary labeling schemes of 8-PAM, as listed in Table 1. In the PAS system, the left bit at each symbol point is the sign bit indicating positive or negative sign information, and the following middle and right bits are amplitude bits indicating the amplitude information. In addition, the PAS system restricts that the amplitude bits are symmetrically arranged; in other words, symbol points with the same absolute value have the same amplitude bits. Furthermore, binary labeling schemes that invert all bits or only amplitude bits are treated as the same scheme. Based on the above restrictions, there are only three kinds of labeling schemes of 8-PAM for the PAS system, as listed in Table 1.

Binary reflected Gray code (BRGC) has 1 bit for the Hamming distance, and it is a well-known binary labeling scheme that is suitable for optimizing BICM systems. In natural-based code (NBC), the amplitude bits decrease when the symbol points increase from +1 to +7 or the symbol points decrease from −1 to −7. Finally, non-natural-based binary code (NNBC) is the only other possible binary labeling scheme for PAS. Under the same signal — noise ratio (SNR), a BICM system employing BRGC has a better BER than BICM systems employing NBC and NNBC. In addition, a BICM system employing NBC has a better BER than a BICM system employing NNBC. This is because a single symbol error event leads to multiple bit errors in the systems employing NBC and NNBC but is easily estimated from the average Hamming distance.

Table 1 All possible binary labeling schemes of 8-PAM.

Symbol point	−7	−5	−3	−1	+1	+3	+5	+7
BRGC	000	001	011	010	110	111	101	100
NBC	000	001	010	011	111	110	101	100
NNBC	000	011	010	001	101	110	111	100

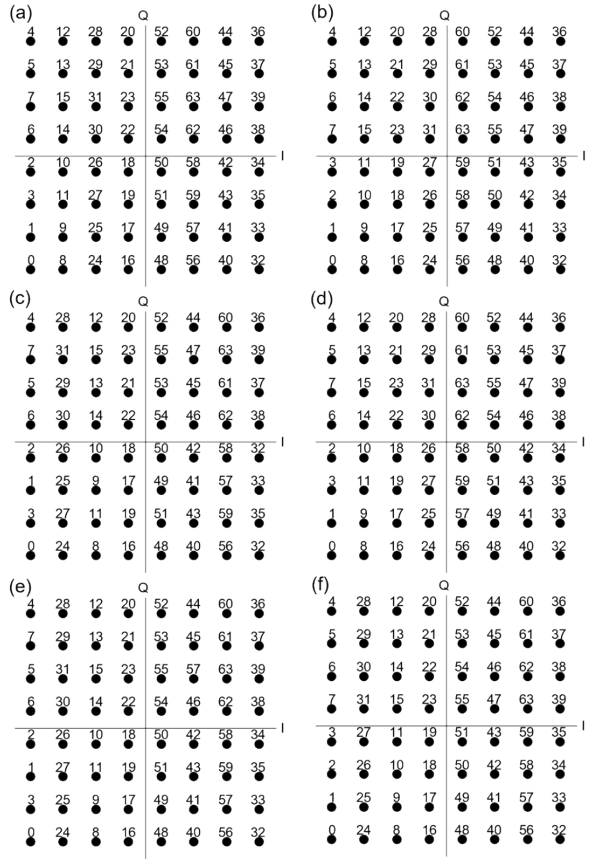


Fig. 2 All possible binary labeling schemes of 64-QAM in decimal: (a) BRGC, (b) NBC, (c) NNBC, (d) BRGC × NBC, (e) BRGC × NNBC, and (f) NBC × NNBC.

The binary labeling of 64-QAM requires combining these binary labeling schemes of 8-PAM on I and Q elements independently. Figure 2 shows all possible binary labeling schemes of 64-QAM in decimal. Note that the labeling scheme that exchanged labeling schemes on I and Q elements from the ones shown in Figs. 2(d)–(f) is treated as the same labeling scheme in this paper. BRGC, NBC, and NNBC combine the same binary labeling scheme on the I and Q elements. In contrast, BRGC × NBC, BRGC × NNBC, and NBC × NNBC combine different binary labeling schemes on the I and Q elements.

4. Numerical Simulation

4.1 Simulation Setup

We used common simulation setups for the AWGN and optical fiber channels. For PAS, we used the 64-QAM format (i.e., $M = 8$ and $m = 3$). The shaping method was ESS [16]. We set the length of input bit sequence and amplitude sequence for ESS to $k_s = 48$ and $n_s = 32$, respectively, by selecting the bounded-energy of 280 in an amplitude sequence. Then, the probability of each amplitude was $(1, 3, 5, 7) = (0.505, 0.330, 0.134, 0.031)$. The entropy of the amplitude was $H(A) = 1.568$ [bits/amplitude]. The rate loss was $R_{loss} = H(A) - k_s/n_s = 1.568 - 1.5 = 0.068$ [bits/amplitude]. The information rate of the PAS with BICM-ID system was $R = 2(k_s/n_s + 1 - m(1 - R_c)) = 3$ [bits/symbol]. For the FEC encoder and decoder, we used an LDPC code defined according to the Digital Video Broadcasting — Satellite — Second Generation (DVB-S2) standard [24]. The code word length was 64,800 bits, and the code rate was 2/3. For BICM-ID, we set the number of inner iterations to 20. Note that the inner iterations represented iterative decoding in the LDPC decoder. We set the number of outer iterations to 5. Note that the outer iteration represented the feedback from the LDPC decoder to the demodulator. After five outer iterations, the aggregate number of inner iterations becomes $120 = 20 \times (1 + 5)$. Note that the number of outer iterations indicates the number of feedbacks from the decoder, not the number of decoding.

4.2 Performance over the AWGN Channel

We evaluated the dependence of the system performance on the binary labeling scheme. Figure 3 shows the BER performance with BRGC and NBC over the AWGN channel. The system with zero outer iterations (i.e., the BICM system) was used as a baseline. BRGC did not improve the BER performance regardless of the number of outer iterations. In contrast, NBC improved the BER performance, and the performance improved with the number of outer iterations. The SNR requirement for a BER of 10^{-5} was 0.5 dB lower when NBC with five outer iterations was employed than when BRGC was employed. These results indicate that the system performance (i.e., improvement in the SNR and required number of outer iterations) depends on the binary labeling scheme. This agrees with previous results in the literature [6], [7], [20].

The reason of the performance improvement on NBC is improving the reliability of the demodulator and is understandable from (1) and (3). In (1), the $p\left(x \middle| \boldsymbol{L}_i^a\right)$ terms indicate the conditional probability of the symbol given a-priori information. It weights the contribution of the symbol, i.e., $p(x)p\left(y_i \middle| x\right)$, for a specific bit detection. The higher contribution symbols have binary labels closer to a-priori information, as shown in (3). These symbols are neighboring on the constellation labeled by BRGC because the average

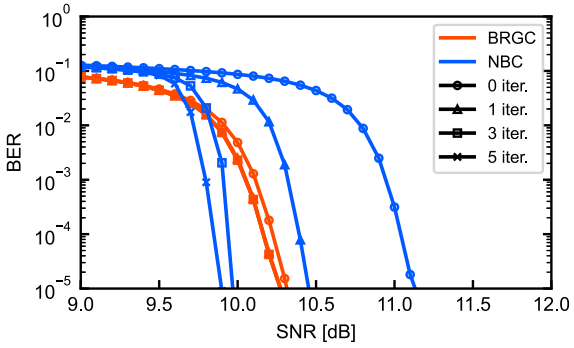


Fig. 3 BER performances over the AWGN channel of PAS with BICM-ID systems employing BRGC and NBC with variable outer iterations.

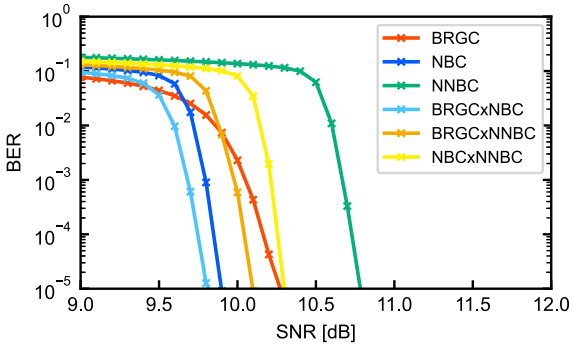


Fig. 4 BER performances over the AWGN channel of PAS with BICM-ID systems employing different binary labeling schemes with five outer iterations.

Hamming distance is 1. However, in the other cases, e.g., NBC, the symbols are not neighboring because the average Hamming distance is greater than 1. This implies that when high a-priori information is given with some outer iterations, the minimum Euclidean distance between the high contribution symbols in NBC is greater than in BRGC. Consequently, the system employing NBC outperforms the one employing BRGC, as shown in Fig. 3, since the longer minimum Euclidean distance leads to the higher reliability of the demodulator.

Based on the observed dependence on the binary labeling scheme, we investigated the most suitable scheme according to the BER with five outer iterations. Figure 4 shows the BER versus SNR of the systems employing all possible binary labeling schemes with five outer iterations. The BER performances of labeling schemes combining different binary labeling schemes on the I and Q elements are poorly correlated with that of the labeling scheme combined on I or Q element. This is because a small change in the binary label significantly changes the BICM-ID gain. Therefore, it is valuable that the evaluation for all possible binary labeling schemes. As a result, the system employing BRGC × NBC had the best BER performance with an SNR gain of 0.55 dB at a BER of 10^{-5}.

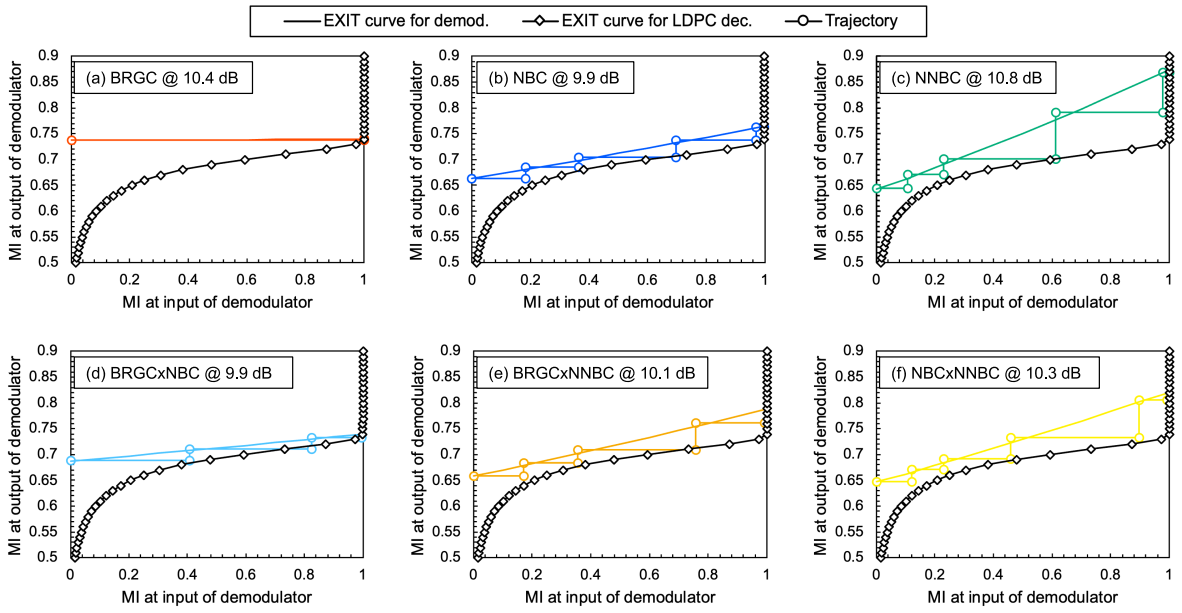


Fig. 5 EXIT charts and mutual information trajectory obtained from the actual simulation results for all possible labeling schemes of 64QAM: (a) BRGC, (b) NBC, (c) NNBC, (d) BRGC × NBC, (e) BRGC × NNBC, (f) NBC × NNBC. SNR at each EXIT chart is the minimum one to achieve BER of 10^{-5} with five outer iterations, as shown in Fig. 4.

4.3 EXIT Chart Analysis

In this subsection, we address three topics based on EXIT chart analysis. One of them is the applicability of using EXIT chart analysis for PAS with BICM-ID system. The second is the most suitable labeling scheme from the EXIT chart and mutual information trajectory obtained in the actual simulation. The last is a guideline for predicting the labeling scheme that provides the high coding gain.

EXIT chart is a tool for analyzing the iterative coding scheme, e.g., BICM-ID. It consists of some EXIT curves which are expressed as bitwise mutual information (MI) (sometimes called normalized generalized MI [15]), $0 \le I \le 1$. Generally, the EXIT chart for BICM-ID system deals with two EXIT curves for the decoder and demodulator. Both curves show the relationship between a-priori information and extrinsic information for the decoder and demodulator. The intersection of EXIT curves indicates the convergence point which is the maximum performance with an arbitrary number of outer iterations. In addition, the tunnel between EXIT curves shows the iterative exchange of extrinsic information between the decoder and demodulator.

Figure 5 shows the EXIT chart and MI trajectory for all labeling schemes. The SNR set at each EXIT chart to achieve BER of 10^{-5} with five outer iterations, as illustrated in Fig. 4. In addition, bitwise MI for the trajectory is calculated by [25]

$$I \approx 1 - \frac{1}{nm} \sum_{i=1}^{n/2} \sum_{j=1}^{2m} \log_2 \left[1 + \exp\left((-1)^{c_i(j)} L_i(j) \right) \right], \quad (4)$$

where $L_i(j)$ is a-priori LLR or extrinsic LLR of the demodulator. This function is valid for non-Gaussian or unknown distributions of LLR when using a sufficiently large number of samples [25]. Note that a-priori information of the demodulator (decoder) is equal to extrinsic information of the decoder (demodulator) because extrinsic information of the decoder (demodulator) is fed into the demodulator (decoder) as a-priori information.

Here, we confirm the applicability of the EXIT chart analysis to PAS with BICM-ID system from the agreeability between the EXIT chart and MI trajectory of the actual simulation. From Fig. 5, EXIT charts and the trajectories were almost perfectly matched, except for Fig. 5(c). As shown in Fig. 5(c), the system employing NNBC had a slight difference between the EXIT chart and the trajectory at the high a-priori information region. This is mainly because the proposed system interleaves only the information bits of the LDPC code. Other bits (parity bits of LDPC code) are regularly placed in the sign bits forming a QAM symbol so the independence between the decoder and demodulator is slightly broken in the actual system. It can be considered that this impact depends on the code rate which represents the ratio of the length of interleaved information bits to the code word. For example, when the code rate is low, the ratio of information bits to be interleaved decreases, and the effect is larger. The code rate of 2/3 used in this paper is the worst condition because it is the lowest one in 64-QAM with PAS. However, the mismatching was only in the high a-priori information region of the system employing NNBC and the impact of the effect was small. From this fact, it can be concluded that using the EXIT chart as an analysis tool for the PAS with BICM-ID is valid under most conditions.

Next, we evaluate the most suitable labeling schemes in

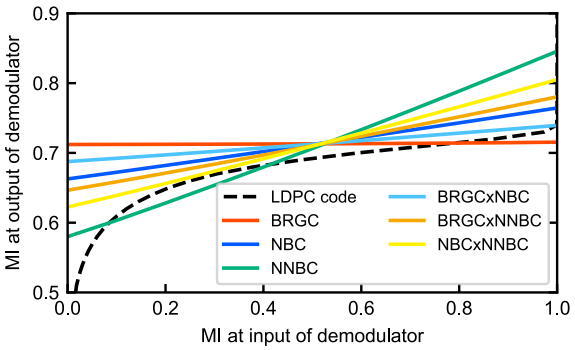


Fig. 6 EXIT curves of the demodulator with all possible binary labeling schemes at an SNR of 9.9 dB and the EXIT curve of the decoder with the LDPC code at a code rate of 2/3.

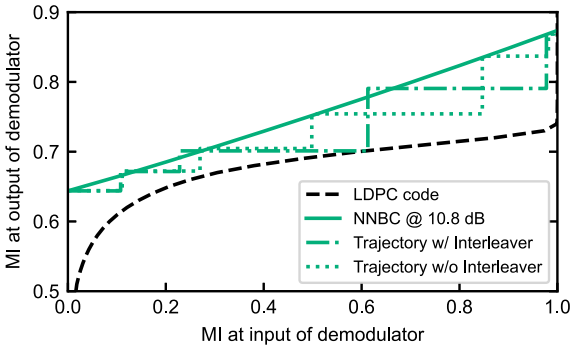


Fig. 7 EXIT curves of the demodulator employing NNBC at an SNR of 10.8 dB and the EXIT curve of the decoder with the LDPC code at a code rate of 2/3.

this simulation setup. Since SNRs in Figs. 5(a)–(f) are the minimum ones for each labeling scheme to achieve BER of 10^{-5}, NBC and BRGC × NBC provide the highest performance of all possible labeling schemes of 64-QAM. Moreover, their trajectories in Figs. 5(b) and d indicate the difference between NBC and BRGC × NBC from the perspective of calculation cost. The system employing NBC required four outer iterations to achieve a-priori information of 1. However, when employing BRGC × NBC, the system required only two outer iterations. As the number of outer iterations directly relates to the calculation cost, this fact shows that BRGC × NBC is the most suitable labeling scheme in this simulation setup.

Finally, we introduce a guideline for predicting the labeling scheme that provides the high coding gain based on the EXIT chart. Figure 6 shows the EXIT chart at SNR of 9.9 dB which is required to achieve BER of 10^{-5} when employing the most suitable labeling in Fig. 4. It includes one EXIT curve for the decoder and six EXIT curves for the demodulator employing all possible binary labeling schemes of 64-QAM. As can be seen in Fig. 6, NBC, BRGC × NBC, and BRGC × NNBC were suitable labeling schemes for this SNR condition, focusing on the convergence points and the slope of the EXIT curves. This is because, in their labeling schemes, the slopes of the EXIT curve for the decoder and demodulator are similar. In this condition, two EXIT curves open the tunnel for the extrinsic information exchange at lower SNR. Note that the narrower tunnel requires more outer iterations to achieve target BER.

Concretely, binary labeling schemes with a steep slope (e.g. NNBC and NBC × NNBC) have the convergence point at low a-priori information for the demodulator. In addition, a binary labeling scheme with a gradual slope (e.g. BRGC) has the convergence point at high a-priori information for the demodulator but less than 1. In the system employing such labeling schemes, the BER of 10^{-5} cannot be achieved at 9.9 dB with an arbitrary number of outer iterations. However, binary labeling schemes with a similar slope of EXIT curves of the demodulator to one of EXIT curves of the decoder (e.g. NBC, BRGC × NBC, and BRGC × NNBC) have a convergence point at 1. Thus, these labeling schemes are suitable at this SNR condition, and the system employing them achieves the BER of 10^{-5} with an arbitrary number of outer iterations. These results also indicate that the BICM-ID gain is not directly related to the coding gain. Alamri et al. [27] noted that a binary labeling scheme with a steeper slope yields a higher BICM-ID gain. While this agrees with our results, these binary labeling schemes had a convergence point at less than 1 at lower SNR conditions and did not lead to a high coding gain.

4.4 Effect of Interleaver

To confirm the effect of random interleavers, we compared the number of required outer iterations according to the EXIT chart and actual simulation results as shown in Fig. 7.

Figure 7 includes two trajectories from the system with and without interleavers. They are the average of 100 times simulations. In addition, the EXIT curves of the decoder and the demodulator are the same as in Fig. 5(c). Ideally, the trajectory passed through the tunnel with the edge of each step touching both EXIT curves. However, the trajectory from the system without interleavers showed mismatching, particularly with high a-priori information. Compared with the trajectory from the system without interleaver, the one from the system with interleaver had better agreement with the EXIT chart. This result shows that using interleavers improves the agreeability between EXIT chart and MI trajectory. The reason of improving the agreeability is that it enhances the independence between the decoder and demodulator, which is necessary to maximize their correction performance. To further improve this agreement, there is a way to increase the length of interleaver. This method can be achieved by interleaving across multiple LDPC code words for PAS with BICM-ID. However, this increases calculation complexity and latency, so it is not addressed in this paper. Note that the method cannot improve the fundamental reason as mentioned in the discussion for Fig. 5.

To analyze the effect of interleaver, Fig. 8 shows an example of the input LLR sequences of the decoder in zero, two, and four outer iterations at SNR of 10.8 dB. Note that the signs of some LLR, that correspond to “0” in binary bit,

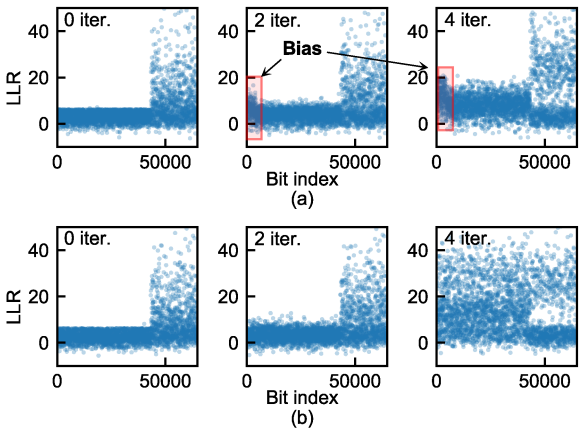


Fig. 8 Example of input LLR sequences of the FEC decoder in zero, two, and four outer iterations at an SNR of 10.8 dB. The sign of some LLR, that correspond to "0" in binary bit, is inverted. PAS with BICM-ID systems (a) without and (b) with random bit interleavers employ NNBC as the binary labeling.

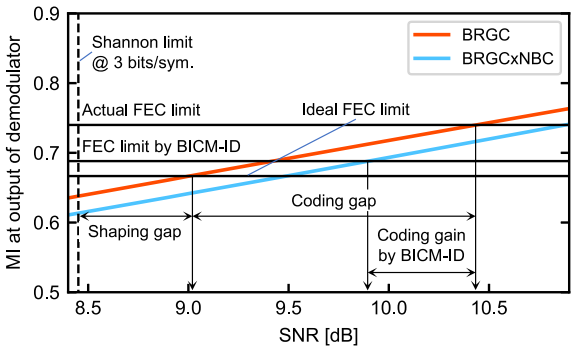


Fig. 9 Bitwise MI at the demodulator output of two labeling schemes, BRGC and BRGC × NBC, as a function of SNR.

are inverted. This implies that the positive value indicates a correct hard decision, and the negative one indicates an incorrect hard decision. In Fig. 8, these systems with and without interleavers employ NNBC as the binary labeling scheme. In the system with interleavers, the decoder inputs at zero, two, and four outer iterations had a uniform distribution, while in the system without interleavers, accuracy biases occurred in the decoder input at two, and four outer iterations. This is because, in the system without interleavers, the correlation between the decoder and demodulator is strong and their error correction capabilities become similar as a-priori information increases. Therefore, the bias strongly depends on the output characteristics of the LDPC decoder, which is determined by the degree distribution of the LDPC code [24], even though it is in the decoder input. In such a situation, the quality of some bits increases while the quality of other bits remains low after many outer iterations, as shown in Fig. 8(a). Therefore, the decoder and demodulator provided lower extrinsic information than expected in the EXIT chart in the system without interleavers, as shown in Fig. 7. On the other hand, the system with random bit interleavers kept uncorrelated and effectively suppresses the bias, as illustrated in Fig. 8(b).

In the PAS with BICM-ID system, with or without interleavers, the decoder may be affected by the performance difference between the bits forming one symbol, called unequal error protection (UEP) property. The reason for this is that the PAS system allocates the parity bits of the LDPC code into the sign bits of the QAM symbol, as mentioned in Sect. 2.1. In the case of such arbitrary bit allocation in the codeword, the decoder is more susceptible to UEP property, then the decoding performance is improved or degraded. The UEP property in the PAS system depends on the binary labeling scheme and the probability of symbol occurrence set by the DM. This paper deals with only a kind of probability distribution so it is not sufficient for a detailed evaluation of the effect of UEP property. Therefore, we leave this topic for future research.

4.5 Bitwise Mutual Information Analysis

By using the bitwise MI at the demodulator output, we evaluate the SNR penalties (e.g., shaping gap and coding gap) to Shannon limit and the coding gain from BICM-ID. Figure 9 shows the MIs of BRGC and BRGC × NBC as a function of SNR, together with three FEC limits and Shannon limit at 3 [bit/symbol]. Ideal FEC limit is the code rate of FEC code which is the required MI to achieve less than 10^{-5} after decoding if the FEC code has ideal performance. Actual FEC limit is the MI required to achieve less than 10^{-5} after decoding without any outer iteration, seen in Fig. 6 as a-priori information for the decoder at the extrinsic information for the decoder of 1. On the other hand, FEC limit by BICM-ID is the required MI for BRGC × NBC at the initial demodulation to achieve less than 10^{-5} after five outer iterations.

We can estimate the SNR penalties due to shaping and coding gap from the MI of BRGC. First, we estimated the SNR penalty due to shaping gap was about 0.55 dB by comparing the SNR at Shannon limit and at ideal FEC limit. The shaping gap depends on the shaping method and the length of input and output block for DM. To further reduce shaping gap, we use longer length although it requires higher calculation cost. Second, the SNR penalty due to coding gap was about 1.4 dB by comparing the SNR at ideal FEC limit with the SNR at FEC limit. This coding gap depends on the error correction capability of FEC code in BICM system. By applying the BICM-ID with suitable labeling schemes and some outer iteration, we can reduce this SNR penalty. Therefore, the reduction is coding gain by BICM-ID.

We evaluate the coding gain by BICM-ID. Focusing on MI, BRGC × NBC was inferior to BRGC at all SNRs in Fig. 8. However, BRGC × NBC can achieve less than 10^{-5} at FEC limit by BICM-ID, resulting the coding gain is about 0.5 dB. This coding gain is offered by relaxing the required MI at initial demodulation by five outer iterations in BICM-ID.

Table 2 Parameters for the single-channel transmission over optical fiber system.

Parameter	Value
Modulation	64-QAM
Number of polarizations	2
Wavelength	1550 [nm]
Symbol rate	32 [GBaud]
Number of channels	1
Channel span	50 [GHz]
Attenuation	0.2 [dB/km]
Dispersion parameter	16 [ps nm^{-1} km^{-1}]
Nonlinear coefficient	1.27 [W^{-1} km^{-1}]
Noise figure	5 [dB]
Span length	100 [km]
Number of spans	30–60
Total transmission distance	3000–6000 [km]

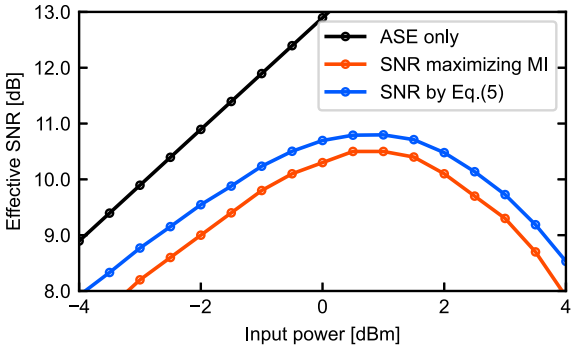


Fig. 10 Effective SNR estimation at 4100 km: only ASE, SNR maximizing the MI at the demodulator, and estimation by (5).

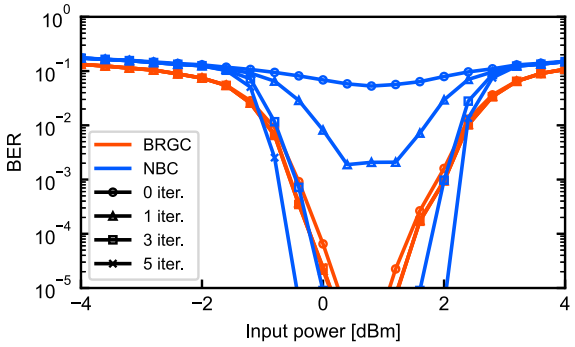


Fig. 11 BER as a function of the input power for the PAS with BICM-ID system employing BRGC and NBC over a transmission distance of 4100 km.

4.6 Optical Fiber Simulation Setup and Effective SNR

Table 2 lists the parameters used to simulate the single-channel transmission over the optical fiber system. On the transmitter side, 64-QAM symbols were generated by using the modulation process introduced in Sect. 2 and were separated evenly and combined on each polarization to generate dual-polarized (DP)-64-QAM symbols. After the DP-64-QAM symbols were up-sampled to eight samples/symbol, pulse shaping was performed with a raised cosine filter and roll-off factor of 0.1. The transmission link consisted of a multi-span standard single-mode fiber. At the end of each span, an erbium-doped fiber amplifier (EDFA) amplified the signal to compensate for fiber loss. This propagation was based on calculating the Manakov equation by the split-step Fourier method [28, Sect. 2.4.1]. After the propagation of all spans, we added the total amplified spontaneous emission (ASE) noise for the number of all EDFAs. On the receiver side, the signal was filtered by the optical filter and was down-sampled to two samples/symbol. Electronic dispersion compensation and an adaptive filter based on the decision-directed least mean square were performed for pulse equalization [1]. Then, the demodulator calculated the LLRs.

Since we assumed (2) as the law of the optical fiber channel, it is needed to estimate SNR after transmission which includes ASE noise and the degradation of the Kerr effect, called effective SNR. We estimated the effective SNR as follows [29]:

$$\mathrm{SNR}_{\mathrm{eff}} = \frac{Var\left[X'\right]}{Var\left[Y - X'\right]} \tag{5}$$

where $Var[\cdot]$ is the variance and X' are the transmitted symbols adjusted to represent centroids of the corresponding received symbols Y.

We confirmed the estimation accuracy of the effective SNR by comparing it to the SNRs that emaximize the bitwise MI. Figure 10 shows the obtained SNR as a function of the input power at a transmission distance of 4100 km for a PAS system employing BRGC. The black dotted curve only considers ASE noise. The red dotted curve shows the SNR that maximizes the MI at the demodulator. This SNR was heuristically selected by incrementing SNR in steps of 0.1 dB. The blue dotted curve shows the SNR estimated by (5). Although the SNR considering only ASE increased with the input power, the effective SNR estimated by (5) and SNR maximizing the MI degraded when the input power was greater than 1 dBm because of the optical Kerr effect. The effective SNR estimated by (5) almost agreed with the SNR that maximized the MI. Thus, we used (5) to estimate the effective SNR in the following optical fiber simulations.

4.7 Performance over the Optical Fiber Channel

We performed numerical simulations to evaluate the system performance for single-channel optical fiber transmission and to consider the degradation due to the optical Kerr effect. Figure 11 shows the BER as a function of the input power of the PAS with BICM-ID system employing BRGC and NBC over a transmission distance of 4100 km. All the BER curves degraded when the input power was increased to more than 1 dBm because of the optical Kerr effect. This indicates that the optimal input power was 1 dBm. In the linear regime with a low input power (i.e., $<$ 1 dBm), the system employing NBC improved the BER performance with more outer iterations, similar to the results over the AWGN channel (see Fig. 3). In the nonlinear regime with a high input power (i.e., $>$ 1 dBm), the system employing NBC showed a greater tolerance of nonlinearity and extended the input

power achieved at a BER of 10^{-5} compared to the system employing BRGC. Thus, the iterative decoding improves the system performance in both regimes.

If the assumption that the law of the optical fiber channel is (2) is not valid, there should be differences in the characteristics of each symbol since the Kerr effect depends on the intensity-dependent. In this case, the degradation due to the Kerr effect depends on the binary labeling schemes. To compare the system performances with all possible binary labeling schemes, Fig. 12 plots the BER as a function of the input power with five outer iterations. The most suitable binary labeling scheme was BRGC × NBC, which is consistent with the results over the AWGN channel. Moreover, the ordering of the binary labeling schemes by BER performance matched the results over the AWGN channel. This consistency was observed in both low and high optical power regions, where the optical Kerr effect has minimal and significant impact respectively. From these results, we can conclude that the assumption that the law of the optical fiber channel is (2) is valid in the proposed system. Further evaluation will be presented in Sect. 4.8.

Because the SNR gain over the AWGN channel corresponds to the transmission distance, we evaluated the transmission distance of the system when employing BRGC, NBC, and BRGC × NBC with five outer iterations. Figure 13 shows the simulation results at an input power of 1 dBm. The systems employing NBC and BRGC × NBC outperformed the system employing BRGC at a BER of 10^{-5}. The transmission distance gains were 300 and 400 km, respectively.

4.8 Degradation due to the Optical Kerr Effect for the Bit Position

To clarify the degradation due to the optical Kerr effect for the bit position, we compare the back-to-back (B2B) performance and the performance after transmission. In B2B performance, we only load the noise and eliminate the transmission based on the split-step Fourie method from the system shown in Sect. 4.6. Figure 14 shows the result of BRGC. It includes three bitwise MIs regarding the bit position. $c_i(1)$ and $c_i(4)$, $c_i(2)$ and $c_i(5)$, and $c_i(3)$ and $c_i(6)$ indicate the left, middle, and rigth bit, respectively, in the

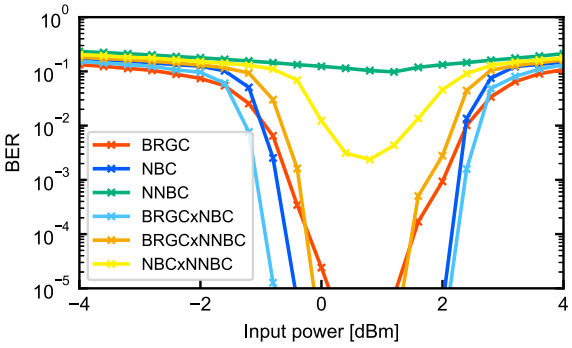


Fig. 12 BER as a function of input power of the PAS with BICM-ID system for all possible binary labeling schemes with five outer iterations and a transmission distance of 4100 km.

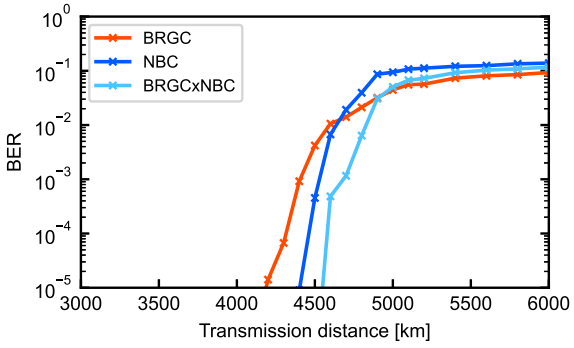


Fig. 13 BER as a function of the transmission distance of the PAS with BICM-ID system employing BRGC, NBC, and BRGC × NBC with five outer iterations at an input power of 1 dBm.

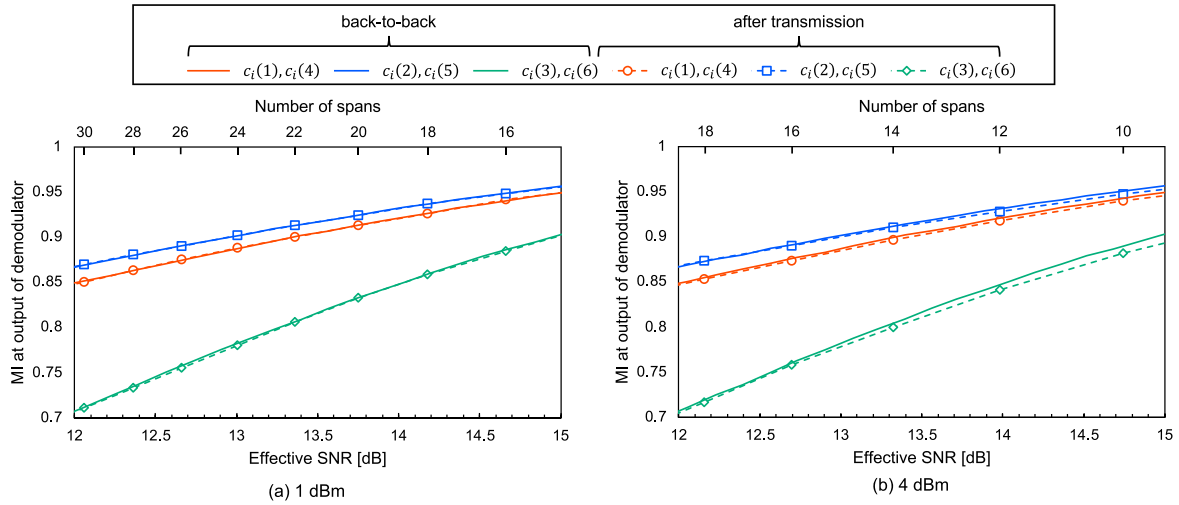


Fig. 14 Bitwise MI performance of each bit position at B2B and after transmission. The input power is set to (a) 1 dBm, the optimal power, and (b) 4 dBm, higher power.

PAM symbol shown in Table 1. We set the input power to 1 dBm, which is the optimal power, in Fig. 14(a), and to 4 dBm, higher power than the optimal power, in Fig. 14(b). These figures have two x-axes: effective SNR and number of spans. Effective SNR at each plot is calculated after the transmission of each number of spans.

We can see in Fig. 14(a) that three MIs after transmission were almost the same as ones of B2B. This implies that the intensity-dependent degradation of the optical Kerr effect is small at the optimal power. In the case of higher power shown in Fig. 14(b), the MIs after transmission had a slight degradation compared to ones of B2B at 10–12 spans. Also, the degradation differed to the bit position and the worst position was $c_i(3)$ and $c_i(6)$. This implies that higher-intensity symbols have poorer performance than lower-intensity ones since the bit position has a bit error when symbol error occurs between higher-intensity symbols. Furthermore, the increase in the number of spans reduced the MI difference between the B2B performance and the performance after transmission, as shown in Fig. 14(b). This is because pulse spreading of wavelength dispersion reduces the difference in symbol intensity. Thus, we can conclude that, in long transmission distance and at the optimal power, the degradation due to the optical Kerr effect can be treated as SNR degradation. This conclusion allows us to determine the suitable binary labeling scheme in the AWGN channel instead of the optical fiber channel, which requires a massive calculation cost, for optical fiber communication.

5. Conclusion

We evaluated the performances of the PAS and BICM-ID system with all possible binary labeling schemes of 64-QAM over the AWGN and optical fiber channels. The simulation results showed that the most suitable binary labeling scheme (BRGC × NBC) was better than using BRGC over both channels. The SNR gain over the AWGN channel was about 0.55 dB, and the transmission distance gain over the optical fiber channel was about 400 km. These results indicated that the PAS with BICM-ID system has a significant gain over the optical fiber channel as well as the AWGN channel. This is because BICM-ID offers a greater tolerance of nonlinearity. Moreover, it is slight that the dependence between the binary labeling scheme and degradation due to the optical Kerr effect at long transmission distances and low input power regions. It leads that the most suitable binary labeling scheme can be determined by simulations of the AWGN channel or EXIT chart rather than the optical fiber channel requiring a massive calculation cost.

We used an EXIT chart to analyze the suitability of binary labeling schemes and the effect of random bit interleavers. The results showed that a binary labeling scheme is suitable when the slope for the EXIT curve of the demodulator is close to the slope for the EXIT curve of the decoder. This finding is valid for system design guidelines, e.g., selecting a suitable labeling scheme from the EXIT chart. In addition, inserting random bit interleavers mitigated the performance degradation that occurs during iterative decoding of BICM-ID. The most suitable binary labeling scheme shown in this paper is not general. This is because it depends on the FEC codes and the probability distribution defined in DM. However, it can be expected from the EXIT chart analysis based on the guidelines shown in Sect. 4.3.

We used bitwise mutual information for the system employing BRGC and NBC to evaluate SNR penalty and gain offered by iterative decoding of BICM-ID. The evaluation showed that the SNR penalties due to shaping gap and coding gap were about 0.5 dB and 1.4 dB respectively. Moreover, the gain by BICM-ID is coding gain which was about 0.5 dB.

References

[1] K. Kikuchi, "Fundamentals of coherent optical fiber communications," J. Lightwave Technol., vol.34, no.1, pp.157–179, Jan. 2016. doi: 10.1109/JLT.2015.2463719.
[2] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," IEEE Trans. Commun., vol.40, no.5, pp.873–884, May 1992. doi: 10.1109/26.141453.
[3] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulationion," IEEE Trans. Inf. Theory, vol.44, no.3, pp.927–946, May 1998. doi: 10.1109/18.669123.
[4] X. Li and J.A. Ritcey, "Bit-interleaved coded modulation with iterative decoding," IEEE Commun. Lett., vol.1, no.6, pp.169–171, Nov. 1997. doi: 10.1109/4234.649929.
[5] S. ten Brink, J. Speidel, and R.-H. Yan, "Iterative demapping and decoding for multilevel modulation," IEEE GLOBECOM, 1998 (Cat. NO. 98CH36250), Nov. 1998. doi: 10.1109/glocom.1998.775793.
[6] S. Benmahmoud and A. Djebbari, "A new improved symbol mapper/8-ary constellation for BICM-ID," Wirel. Eng. Technol., vol.04, no.2, pp.65–70, 2013. doi: 10.4236/wet.2013.42010.
[7] F. Schreckenbach, N. Gortz, J. Hagenauer, and G. Bauch, "Optimized symbol mappings for bit-interleaved coded modulation with iterative decoding," GLOBECOM'03. IEEE Global Telecommun. Conference (IEEE, Cat. no.03CH37489), Dec. 2003. doi: 10.1109/GLOCOM.2003.1258849.
[8] Z. Yang, Q. Xie, K. Peng, and J. Song, "Labeling optimization for BICM-ID systems," IEEE Commun. Lett., vol.14, no.11, pp.1047–1049, Nov. 2010. doi: 10.1109/LCOMM.2010.093010.101049.
[9] M.C. Valenti, R. Doppalapudi, and D. Torrieri, "A genetic algorithm for designing constellations with low error floors," 42nd Annual Conference on Information Sciences and Systems, March 2008. doi: 10.1109/CISS.2008.4558693.
[10] N.S. Loghin, J. Zollner, B. Mouhouche, D. Ansorregui, J. Kim, and S.-I. Park, "Non-uniform constellations for ATSC 3.0," IEEE Trans. Broadcast., vol.62, no.1, pp.197–203, March 2016. doi: 10.1109/TBC.2016.2518620.
[11] B. Chen, C. Okonkwo, H. Hafermann, and A. Alvarado, "Increasing achievable information rates via geometric shaping," Eur. Conference on Optical Communication (ECOC), Sept. 2018. doi: 10.1109/ECOC.2018.8535358.
[12] B. Chen, C. Okonkwo, D. Lavery, and A. Alvarado, "Geometrically-shaped 64-point constellations via achievable information rates," 20th International Conference on Transparent Optical Networks (ICTON), July 2018. doi: 10.1109/ICTON.2018.8473932.
[13] B. Chen, Y. Lei, Z. Liang, W. Ling, X. Xue, and A. Alvarado, "Geometrically-shaped multi-dimensional modulation formats in coherent optical transmission systems," arXiv preprint, Aug. 2022. [Online]. Available at: https://arxiv.org/abs/2207.01152
[14] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," IEEE Trans. Commun., vol.63, no.12, pp.4651–4665, Dec. 2015. doi: 10.1109/TCOMM.2015.2494016.

[15] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," J. Lightwave Technol., vol.37, no.6, pp.1579–1589, March 2019. doi: 10.1109/JLT.2019.2895065.
[16] A. Amari, S. Goossens, Y.C. Gültekin, O. Vassilieva, I. Kim, T. Ikeuchi, C.M. Okonkwo, F.M. Willems, and A. Alvarado, "Introducing enumerative sphere shaping for optical communication systems with short blocklengths," J. Lightwave Technol., vol.37, no.23, pp.5926–5936, Dec. 2019. doi: 10.1109/JLT.2019.2943938.
[17] Y.C. Gültekin, T. Fehenberger, A. Alvarado, and F.M.J. Willems, "Probabilistic shaping for finite blocklengths: Distribution matching and sphere shaping," Entropy (Basel), vol.22, no.5, p.581, May 2020. doi: 10.3390/e22050581.
[18] B. Kien Khoo, S.Y. Le Goff, B.S. Sharif, and C.C. Tsimenidis, "Bit-interleaved coded modulation with iterative decoding using constellation shaping," IEEE Trans. Commun., vol.54, no.9, pp.1517–1520, Sept. 2006. doi: 10.1109/TCOMM.2006.881181.
[19] T. Arafa, W. Sauer-Greff, and R. Urbansky, "Performance of combined constellation shaping and bit interleaved coded modulation with iterative decoding (BICM-ID)," Adv. Radio Sci., vol.9, pp.195–201, Aug. 2011. doi: 10.5194/ars-9-195-2011.
[20] A. Naka, "Performance of probabilistic amplitude shaping with BICM-ID," Electron. Lett., vol.57, no.5, pp.226–228, Jan. 2021. doi: 10.1049/ell2.12093.
[21] M. Komatsu and A. Naka, "Probabilistic amplitude shaping with BICM-ID for all possible labels on 64-QAM," 27th OptoElectronics and Communications Conference (OECC) and International Conference on Photonics in Switching and Computing (PSC), vol.2022, July 2022. doi: 10.23919/OECC/PSC53152.2022.9850181.
[22] Z. Yang, Q. Xie, K. Peng, and Z. Wang, "A novel BICM-ID system approaching shannon-limit at high spectrum efficiency," IEICE Trans. Commun., vol.E94-B, no.3, pp.793–795, March 2011. doi: 10.1587/transcom.E94.B.793.
[23] A. Naka, "Performance comparison of probabilistic amplitude shaping and multidimensional modulation," IEICE Commun. Express, vol.9, no.4, pp.105–110, 2020. doi: 10.1587/comex.2019XBL0156.
[24] "Digital Video Broadcasting (DVB); 2nd generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications (DVB-S2), European Telecommun. Standards Inst. (ETSI) Standard EN," Rev., vol.302 307, p.1.2.1, 2009.
[25] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," IEEE Trans. Commun., vol.49, no.10, pp.1727–1737, 2001. doi: 10.1109/26.957394.
[26] M. El-Hajjar and L. Hanzo, "EXIT charts for system design and analysis," IEEE Commun. Surveys Tuts., vol.16, no.1, pp.127–153, 2014. doi: 10.1109/SURV.2013.050813.00137.
[27] O. Alamri, B. Poupart, M. El-Hajjar, S.-X. Ng, and L. Hanzo, "On multidimensional BICM-ID constellation labelling," IEEE International Conference on Communications, May 2010. doi: 10.1109/ICC.2010.5502739.
[28] G.P. Agrawal, Fiber Optic Communication Systems, 2nd ed., John Wiley & Sons, New York, 1997.
[29] P. Skvortcov, I. Phillips, W. Forysiak, T. Koike-Akino, K. Kojima, K. Parsons, and D.S. Millar, "Nonlinearity tolerant lut-based probabilistic shaping for extended-reach single-span links," IEEE Photon. Technol. Lett., vol.32, no.16, pp.967–970, Aug. 2020. doi: 10.1109/LPT.2020.3006737.

Mamoru Komatsu received a B.E. degree from the Department of Media and Telecommunications Engineering, Ibaraki University, Japan, in 2021. He received a M.E. degree in Electrical and Electronic Systems Engineering, Graduate School of Science and Engineering, Ibaraki University, Japan, in 2023. He is currently pursuing a Ph.D. degree in Society's Infrastructure Systems Science, Graduate School of Science and Engineering, Ibaraki University, Japan. He is a student member of IEICE.

Akira Naka received a B.S. and M.S. in Physics and a Dr. Eng. in Electrical and Electronic Engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1987, 1989, and 2008, respectively. In 1989, he joined the NTT Transmission Systems Laboratories and moved to NTT Network Service Systems Laboratories in 2001, where he researched optical amplifiers and in-line amplifier systems. In 2015, he became an Associate Professor at the Department of Media and Telecommunications Engineering, Ibaraki University. He has been a Professor at the Department of Electrical and Electronic Systems Engineering since 2018. Dr. Naka is an IEEE member and IEICE member.

PAPER

Reduction of Fiber Four-Wave Mixing Generated from Modulated Lights by Inserting Dispersive Elements

Ayano INOUE[†], ***Nonmember***, **Koji IGARASHI**[†], **Shigehiro TAKASAKA**[††], ***and*** **Kyo INOUE**[†a)], ***Members***

SUMMARY Four-wave mixing (FWM) is a crucial impairment factor in optical wavelength-division-multiplexing (WDM) transmission systems over dispersion-shifted fibers. This paper presents an FWM suppression scheme that places dispersive elements (DEs) such as dispersion compensation fibers at optically repeating points in transmission lines. In a DE, the relative phase of the transmitted signal lights and the FWM light generated in the previous spans is shifted. Consequently, the FWM lights generated in each span are summed in random phases and the total FWM power at the end of the transmission lines is reduced from that in straight transmission lines with no DEs. We conduct proof-of-principle experiments to confirm the mechanism of the FWM reduction. Calculation for evaluating the FWM reduction ratio in a WDM transmission system is also presented.

key words: *fiber four-wave mixing, suppression scheme, chromatic dispersion, modulated signal*

1. Introduction

Four-wave mixing (FWM) is a nonlinear optical phenomenon that generates new wavelengths of light from two or three optical signals of different wavelengths [1]. This phenomenon can degrade wavelength-division-multiplexing (WDM) transmission systems, in which the generated FWM lights overlap onto the signal light and serve as noise [2], [3]. FWM is efficiently generated in dispersion-shifted fibers (DSFs) because the phase-matching condition, under which the nonlinear polarization wave and signal waves co-propagate in phase, is easily satisfied in DSFs [4]. Therefore, DSFs are not employed in optical transmission systems because FWM hinders WDM transmission. However, previously installed DSFs are still used in some transmission systems. The performance of such transmission systems would be improved if FWM generation could be reduced.

Conventionally, the use of non-zero dispersion fibers or the dispersion management has been known as a countermeasure against FWM in WDM systems over DSFs. However, non-zero dispersion fibers must be intentionally installed for this scheme. The use of the L band instead of the C band is also effective to mitigate FWM over DSF transmission lines [5]. However, this countermeasure wastes the C band or cannot be applied to the C-band transmission, for which the fiber attenuation is minimum and standard Erbium-doped fiber amplifiers are available.

Manuscript received February 5, 2024.
Manuscript revised May 12, 2024.
Manuscript publicized June 28, 2024.
[†]Osaka University, Suita-shi, 565-0871 Japan.
[††]Furukawa Electric Co. Ltd., Ichihara-shi, 290-8555 Japan.
a) E-mail: kyo@comm.eng.osaka-u.ac.jp
DOI: 10.23919/transcom.2024EBP3026

On the above background, this paper presents a scheme to mitigate FWM in optical repeating transmission systems over DSFs, where WDM lights are positioned in the zero-dispersion wavelength of the DSFs in the worst case. Dispersive elements (DEs) such as dispersion compensation fibers (DCFs) are inserted into transmission lines. Through a DE, the relative phase between the nonlinear polarization wave and the FWM lights generated in the previous spans is shifted. Consequently, the phases of the FWM lights generated in each span are randomized, and the total FWM power is reduced from that in straight transmission lines without DEs. We analyze the FWM generation in transmission lines with DCFs and conduct proof-of-principle experiments. Subsequently, calculations for evaluating FWM reduction in WDM systems are presented.

2. Analysis

2.1 System Model

The transmission system model considered in this study is illustrated in Fig. 1. It is an optically repeating system over DSFs, in which optical amplifiers are placed between the transmitter and receiver at equal intervals. The amplifier gain is set to a value that compensates for the transmission loss of one span. At the output of each amplifier, a DCF is placed as a DE, which introduces a phase shift between the nonlinear polarization wave and the FWM light, as will be described later. WDM signal lights with identical polarization states are assumed to be transmitted over this transmission line.

2.2 FWM Generation

The FWM light at the receiver can be regarded as the sum of FWM lights generated in each span that linearly propagate to the receiver. We denote the amplitude of the FWM light

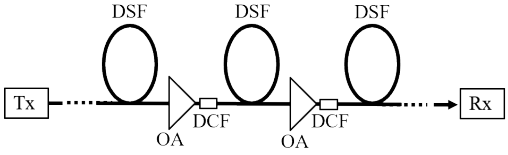


Fig. 1 Transmission system model. Tx: transmitter, Rx: receiver, DSF: dispersion-shifted fiber, OA: optical amplifier, DCF: dispersion compensation fiber.

generated in the kth span at position z by $A_{\mathrm{F}}^{(k)}(z)$. This amplitude at the end of the kth span can be expressed as [6], [7]

$$A_{\mathrm{F}}^{(k)}(z_k + L_0) = id\gamma A_1(z_k) A_2(z_k) A_3^*(z_k) \times \frac{1 - e^{-(\alpha - i\Delta\beta)L_0}}{\alpha - i\Delta\beta} e^{-(\alpha/2 - i\beta_{\mathrm{F}})L_0}, \quad (1)$$

where the FWM light generated at frequency $f_{\mathrm{F}} = f_1 + f_2 - f_3$ is considered, where f_j is the signal-light frequency ($j = 1 - 3$); $A_j(z)$ denotes the signal light amplitude of frequency f_j at position z; γ is the nonlinear coefficient; d is the degeneracy factor, which is 2 for $f_1 \neq f_2$ and 1 for $f_1 = f_2$, $\Delta\beta \equiv \beta_1 + \beta_2 - \beta_3 - \beta_{\mathrm{F}}$ denotes the phase mismatch in a DSF; β_j is the propagation constant for light of frequency f_j; α is the fiber loss coefficient; L_0 is the length of one span; and z_k denotes the position of the input of the kth span.

The signal light at position z_k can be expressed as

$$A_j(z_k) = A_{j0} \exp\left[i(k-1)\left(\beta_j L_0 + b_j L_d\right)\right], \quad (2)$$

where A_{j0} denotes the signal amplitude at $z = 0$, b_j is the propagation constant in a DCF for f_j frequency light, and L_{d} is the DCF length. Substituting Eq. (2) into Eq. (1) yields

$$A_{\mathrm{F}}^{(k)}(z_k + L_0) = id\gamma A_{10} A_{20} A_{30}^* \frac{1 - e^{-(\alpha - i\Delta\beta)L_0}}{\alpha - i\Delta\beta} e^{-(\alpha/2 - i\beta_{\mathrm{F}})L_0} \times \exp\left[i(k-1)\left\{(\beta_{\mathrm{F}} + \Delta\beta) L_0 + (b_{\mathrm{F}} + \Delta b) L_{\mathrm{d}}\right\}\right], \quad (3)$$

where $\Delta b \equiv b_1 + b_2 - b_3 - b_{\mathrm{F}}$ denotes the phase mismatch in a DCF.

The FWM light generated in the kth span propagates linearly to the receiver. Its amplitude at the receiver can be expressed as

$$A_{\mathrm{F}}^{(k)}(\mathrm{end}) = A_{\mathrm{F}}^{(k)}(z_k + L_0) \exp\left[i(N-k)\left(\beta_{\mathrm{F}} L_0 + b_F L_{\mathrm{d}}\right)\right], \quad (4)$$

where N is the total number of spans from the transmitter to the receiver. Substituting Eq. (3) into Eq. (4) and expanding the formula, we obtain the following expression for the FWM light generated in the kth span and reaching the receiver:

$$A_{\mathrm{F}}^{(k)}(\mathrm{end}) = id\gamma A_{10} A_{20} A_{30}^* \frac{1 - e^{(-\alpha + i\Delta\beta)L_0}}{\alpha - i\Delta\beta} e^{(-\alpha/2 + iN\beta_{\mathrm{F}})L_0} \times e^{i(N-1)\beta_{\mathrm{F}} L_{\mathrm{d}}} e^{i(k-1)(\Delta\beta L_0 + \Delta b L_{\mathrm{d}})}. \quad (5)$$

The total FWM light amplitude at the receiver, A_{F}, is expressed by the sum of FWM lights generated in each span and reaching the receiver, as shown below:

$$A_{\mathrm{F}} = \sum_{k=1}^{N} A_{\mathrm{F}}^{(k)}(\mathrm{end}) = id\gamma A_{10} A_{20} A_{30}^* \frac{1 - e^{(-\alpha + i\Delta\beta)L_0}}{\alpha - i\Delta\beta} e^{(-\alpha/2 + iN\beta_{\mathrm{F}})L_0} e^{i(N-1)\beta_{\mathrm{F}} L_{\mathrm{d}}} \times \sum_{k=1}^{N} e^{i(k-1)(\Delta\beta L_0 + \Delta b L_{\mathrm{d}})}. \quad (6)$$

Subsequently, the FWM power at the receiver, P_{F}, is expressed as

$$P_{\mathrm{F}} = |A_{\mathrm{F}}|^2 = (d\gamma)^2 P_0^3 e^{-\alpha L_0} \left|\frac{1 - e^{(-\alpha + i\Delta\beta)L_0}}{\alpha - i\Delta\beta}\right|^2 \cdot \left|\sum_{k=1}^{N} e^{i(k-1)(\Delta\beta L_0 + \Delta b L_{\mathrm{d}})}\right|^2 = (d\gamma)^2 P_0^3 e^{-\alpha L_0} \eta(\Delta\beta) \frac{\sin^2\left[N\left(\Delta\beta L_0 + \Delta b L_{\mathrm{d}}\right)/2\right]}{\sin^2\left[\left(\Delta\beta L_0 + \Delta b L_{\mathrm{d}}\right)/2\right]}, \quad (7)$$

with

$$\eta(\Delta\beta) \equiv \frac{\left(1 - e^{-\alpha L_0}\right)^2 + 4e^{-\alpha L_0} \sin^2\left(\Delta\beta L_0/2\right)}{\alpha^2 + \Delta\beta^2},$$

where $P_0 = |A_{j0}|^2$. For $L_{\mathrm{d}} = 0$, i.e., the case without DCFs, this equation can be rewritten as [8]

$$P_{\mathrm{F}}(L_{\mathrm{d}} = 0) = (d\gamma)^2 P_0^3 e^{-\alpha L_0} \eta(\Delta\beta) \frac{\sin^2\left[N\Delta\beta L_0/2\right]}{\sin^2\left[\Delta\beta L_0/2\right]}. \quad (8)$$

2.3 FWM Reduction

Using Eqs. (7) and (8), the power ratio of the FWM light with and without DCFs, R, can be evaluated as

$$R = \frac{P_{\mathrm{F}}(L_{\mathrm{d}} \neq 0)}{P_{\mathrm{F}}(L_{\mathrm{d}} = 0)} = \frac{\sin^2\left[N\left(\Delta\beta L_0 + \Delta b L_{\mathrm{d}}\right)/2\right]}{\sin^2\left[\left(\Delta\beta L_0 + \Delta b L_{\mathrm{d}}\right)/2\right]} \cdot \frac{\sin^2\left[\Delta\beta L_0/2\right]}{\sin^2\left[N\Delta\beta L_0/2\right]}. \quad (9)$$

For $\Delta\beta L_0 = 0$, i.e., the case where the phase matching condition is satisfied in the DSFs, Eq. (9) can be rewritten as

$$R(\Delta\beta = 0) = \frac{1}{N^2} \cdot \frac{\sin^2\left(N\Delta b L_{\mathrm{d}}/2\right)}{\sin^2\left(\Delta b L_{\mathrm{d}}/2\right)}. \quad (10)$$

This expression indicates that R takes a value from 0 to 1, depending on $\Delta b L_{\mathrm{d}}$. Therefore, the ratio R can be used as an index to represent the FWM power reduction by inserting DCFs.

The reduction ratio, R, depends on $\Delta b L_{\mathrm{d}}$, as shown in Eq. (9), which can be expressed as [9]

$$\Delta b L_{\mathrm{d}} = \frac{2\lambda^2\pi}{c} D_{\mathrm{d}} L_{\mathrm{d}} (f_1 - f_3)(f_2 - f_3), \quad (11)$$

where λ is the light wavelength, c is the light velocity, and D_{d} is the dispersion parameter of a DCF. Equations (10) and (11) indicate that the reduction ratio varies periodically as a function of the signal frequencies, such as $R = 1$ at the signal frequencies satisfying $\Delta b L_{\mathrm{d}}/2 = m\pi$, where m is an integer.

Here, we consider the frequency interval between the neighboring peaks of R for FWM generation at $f_{\mathrm{F}} = 2f_1 - f_3$ from signal lights of frequencies f_1 and f_3. For this partially degenerate FWM process, Eq. (11) can be rewritten as

$$\Delta b L_{\mathrm{d}} = \frac{2\lambda^2\pi}{c} D_{\mathrm{d}} L_{\mathrm{d}} \Delta f^2, \quad (12)$$

where $\Delta f = f_1 - f_3$ is the frequency separation between the two signal lights. Denoting the frequency interval as f_π, the following equation is obtained:

$$\frac{2\lambda^2\pi}{c} D_d L_d \left(\Delta f + f_\pi\right)^2 - \frac{2\lambda^2\pi}{c} D_d L_d \Delta f^2 = 2\pi, \tag{13}$$

from which f_π is expressed as

$$f_\pi \approx \frac{c}{2\lambda^2 D_\mathrm{d} L_\mathrm{d} \Delta f}, \tag{14}$$

where $f_\pi \ll \Delta f$ is assumed. For example, when $D_\mathrm{d} = -160$ ps/km-nm, $L_\mathrm{d} = 2$ km, and $\Delta f = 100$ GHz, f_π is approximately 2 GHz.

In general, the frequency spectrum of a modulated signal light is broadened around the carrier frequency according to signal modulation. Subsequently, the frequency separation Δf is broadened around the mean value of the modulated signal lights. When this frequency broadening is larger than the frequency interval f_π, the FWM reduction ratio R is averaged over the frequency separation as

$$\langle R \rangle = \left\langle \frac{\sin^2\left[N\left(\Delta\beta L_0 + \Delta b L_\mathrm{d}\right)/2\right]}{\sin^2\left[\left(\Delta\beta L_0 + \Delta b L_\mathrm{d}\right)/2\right]} \right\rangle \frac{\sin^2\left[\Delta\beta L_0/2\right]}{\sin^2\left[N\Delta\beta L_0/2\right]}, \tag{15}$$

where $\langle\ \rangle$ represents the average over the frequency separation Δf or $\Delta b L_\mathrm{d}$. The average term in Eq. (15) is calculated as [Appendix]

$$\left\langle \frac{\sin^2\left[N\left(\Delta\beta L_0 + \Delta b L_\mathrm{d}\right)/2\right]}{\sin^2\left[\left(\Delta\beta L_0 + \Delta b L_\mathrm{d}\right)/2\right]} \right\rangle = N. \tag{16}$$

Subsequently, Eq. (15) is rewritten as

$$\langle R \rangle = N \frac{\sin^2\left[\Delta\beta L_0/2\right]}{\sin^2\left[N\Delta\beta L_0/2\right]}. \tag{17}$$

For $\Delta\beta L_0 \approx 0$, this equation becomes

$$\langle R(\Delta\beta L_0 \approx 0) \rangle = \frac{1}{N}. \tag{18}$$

The analysis above indicates that the FWM generation is reduced by a factor of the number of repeating spans in systems where the phase matching condition is satisfied in the DSFs.

3. Experiment

To confirm the mechanism of the FWM reduction described in the previous section, we conducted a proof-of-principle experiment using the setup shown in Fig. 2. Two wavelength lights were combined via an optical coupler, optically amplified, and incident into two cascaded 2.5-km DSFs with an identical zero-dispersion wavelength, between which an optical amplifier and a DCF with a dispersion parameter of −164 ps/km-nm and a length of 2.0 or 0.5 km, were inserted.

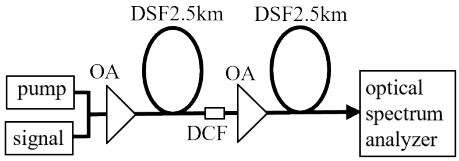


Fig. 2 Experimental setup. DSF: dispersion-shifted fiber, OA: optical amplifier, and DCF: dispersion-compensation fiber.

The DSFs with a length of 2.5 km were employed simply because they were available in our laboratory. The basic ideal of the proposed scheme reducing the FWM power at the end of the transmission line is to randomize the phase relationship between FWM lights generated in each repeating span. Therefore, the FWM power generated in each span, or the span length, is irrelevant to the reduction ratio, and any length of DSFs could be used in the proof-of-principle experiment.

In the above setup, one light was a continuous wave (CW), called "pump light" hereafter, whose wavelength was fixed at the zero-dispersion wavelength of the DSFs. The other light was generated from a wavelength-tunable LD, called "signal light" hereafter, which was CW, quadrature-phase-shift-keying (QPSK) modulated at 12.5 Gbaud, or on-off-keying (OOK) modulated at 12.5 Gbps. The signal wavelength was varied in the longer wavelength side of the pump light. The optical paths from the two light sources to the coupler were constructed using polarization-maintaining fibers, owing to which the two lights were incident to the DSFs in an identical polarization state. The amplifier gain between the DSFs was adjusted such that the optical powers at the inputs of the first and second DSFs were identical as approximately 11 dBm.

The output from the DSFs was incident to an optical spectrum analyzer, and the FWM light power generated in the shorter wavelength side of the pump light was measured. This is because the proposed reduction scheme is effective for FWM satisfying the phase matching condition $\Delta\beta = 0$, as indicated by Eqs. (9) and (10), and the FWM light at the shorter wavelength side satisfied the phase matching condition in the above wavelength allocation.

For the above system condition, the FWM reduction ratio for CW lights can be expressed from Eq. (10) as

$$R\left(N = 2, \Delta\beta L_0 = 0\right) = \cos^2\left(\Delta b L_\mathrm{d}/2\right), \tag{19}$$

where

$$\Delta b = \frac{2\lambda^2\pi}{c} D_\mathrm{d} \left(f_s - f_p\right)^2. \tag{20}$$

In the present experiment, the FWM lights generated in the first and second DSFs interfered with each other at the output end. Therefore, a dual-beam interference pattern would be observed as indicated in Eq. (19). We conducted the following measurement to confirm the formula.

First, a 2.0-km DCF was inserted, and the CW signal was incident. The observed output spectrum is shown in

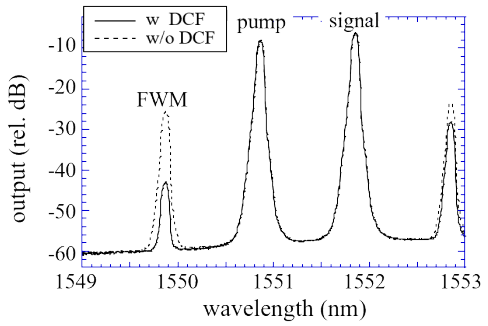


Fig. 3 Output spectrum. DCF length was 2.0 km and signal light was CW. Solid and dashed lines denote outputs with and without DCF, respectively.

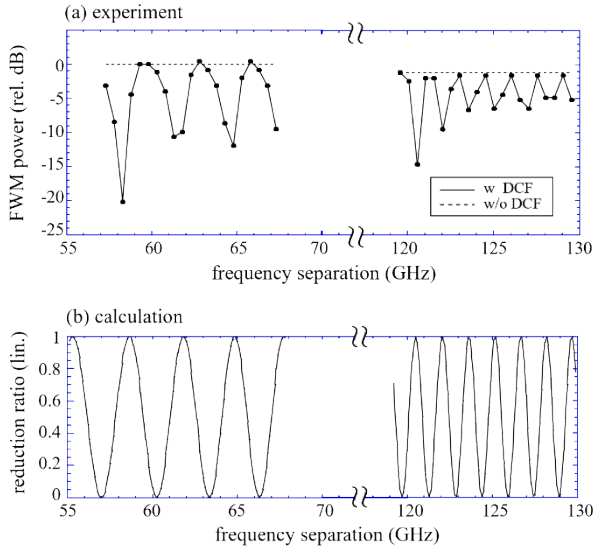


Fig. 4 FWM power as a function of frequency separation between signal and pump lights. DCF length was 2.0 km and signal light was CW.

Fig. 3, where the outputs with and without the DCF are indicated by solid and dashed lines, respectively. In this measurement, the signal light wavelength was carefully chosen at which the FWM power was most effectively reduced. The FWM light in the shorter wavelength side of the pump light was our concern, whose power was significantly reduced by the DCF. On the other hand, the power reduction of the FWM generated in the longer wavelength side was not so significant, because the phase matching condition was not satisfied, i.e., $\Delta\beta \neq 0$, for this FWM light.

The measurement above was performed while the signal-light wavelength was changed, from which the relative FWM power as a function of the frequency separation between the signal and pump lights was plotted. The optical frequency was quoted from the value displayed in the wavelength-tunable LD. The results are shown in Fig. 4(a); additionally, the FWM reduction ratio R calculated using Eq. (19) is shown in Fig. 4(b). As expected from the analytical formula, a periodic output was observed in the experiment, although the extinction ratio was insufficient especially for large frequency separations, and the absolute value

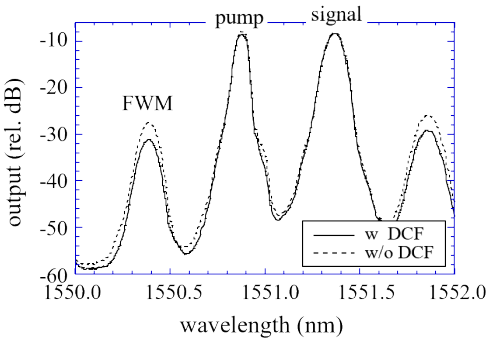


Fig. 5 Output spectrum. DCF length was 2.0 km and signal light was QPSK modulated. Solid and dashed lines denote outputs with and without DCF, respectively.

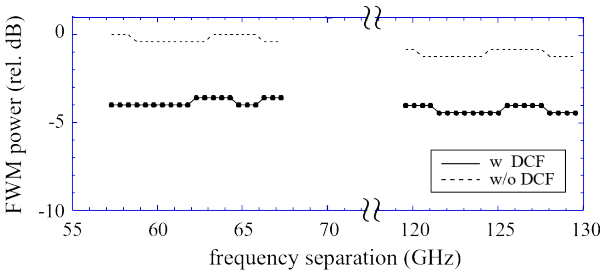


Fig. 6 FWM power as a function of frequency separation between signal and pump lights. DCF length was 2.0 km and signal light was QPSK modulated at 12.5 Gbaud.

of the frequency separation was different between the experimental and calculation results by approximately 1 GHz.

The insufficient extinction ratio was owing to the experimental condition of measuring points in one period being few. In the measurement, we focused on to observe a periodicity or an interference pattern, without caring for the extinction ratio, and thus did not finely change the signal light frequency, being afraid that the experimental conditions varied during the measurement. Subsequently, the number of measuring points were not sufficient to observe a high extinction ratio. The absolute frequency difference between the experiment and calculation might be because the frequency-monitoring system equipped in the wavelength-tunable LD did not have an accuracy within 1 GHz, and the actual frequency differed from the value displayed on the LD module.

Next, the signal light was QPSK modulated instead of CW. The output spectrum and the measured FWM power are shown in Figs. 5 and 6, respectively. Figure 5 indicates that the signal-light spectrum was broadened owing to the QPSK modulation, as was the FWM light. Owing to this spectrum broadening, the FWM power was averaged over the frequency separation and observed to be almost constant at a value that was 3–4 dB less than the FWM power without the DCF, as shown in Fig. 6. This experimental result confirmed the FWM reduction for the modulated signals, as suggested by Eq. (18).

Next, the signal light was OOK modulated instead of

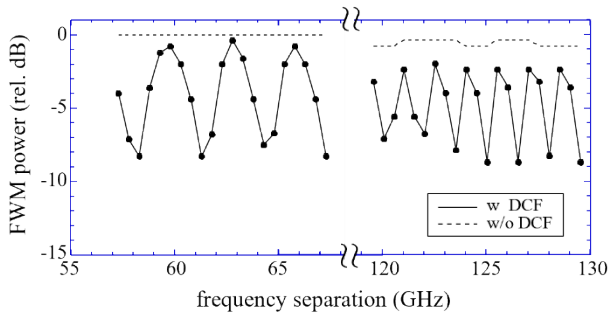


Fig. 7 FWM power as a function of frequency separation between signal and pump lights. DCF length was 2.0 km and signal light was OOK modulated at 12.5 Gbps.

QPSK modulated. The measured FWM power as a function of frequency separation is shown in Fig. 7. In contrast to the result for the QPSK modulated signal shown in Fig. 6, a periodic FWM output was observed even though the modulated light was incident. However, the depth of the periodicity was smaller than that for the CW signal shown in Fig. 4, which is an intermediate property between those of the CW and QPSK signals. This might be because the spectrum of the OOK-modulated light contained a peak at the carrier frequency and was not uniformly broadened, unlike the QPSK signal. Therefore, the OOK signal was in an intermediate state between the CW and QPSK signals in terms of the spectrum, for which the FWM power was partially averaged. As shown in Fig. 7, the discrepancy between the peak FWM level with and without the DCF was larger in the frequency region of 120–130 GHz than in the range of 60–70 GHz. This might be because the frequency period in the former frequency region was narrower and the averaging effect was more effective.

Subsequent to the use of a 2.0-km DCF, we examined a DCF with a length of 0.5 km. The results for the CW signal are shown in Fig. 8, where the measured FWM power and FWM reduction ratio calculated using Eq. (19) are plotted in (a) and (b), respectively. Frequency periodicity was observed, similar to the results for the 2.0-km DCF shown in Fig. 4; however, the period was four times larger than that for the 2.0-km DCF, thus corresponding to the fact that the 0.5-km DCF is four times shorter than the 2.0-km DCF.

Additionally, we examined the 12.5-Gbaud QPSK signal for the system with a 0.5-km DCF; the result is shown in Fig. 9. In contrast to the result of the system with a 2.0-km DCF, periodicity was observed in the frequency region of 40–90 GHz, even when the signal light was QPSK modulated. This is attributed to the insufficient averaging effect arising from the wide frequency period shown in Fig. 8. The experimental results above based on the 0.5-km DCF indicate that the DCF length should be selected appropriately by considering on the frequency separation and signal modulation bandwidth to obtain the averaging effect.

The experimental results above showed the FWM generation characteristics expected from the theoretical analysis in the previous section, confirming the mechanism of FWM reduction by inserting DEs.

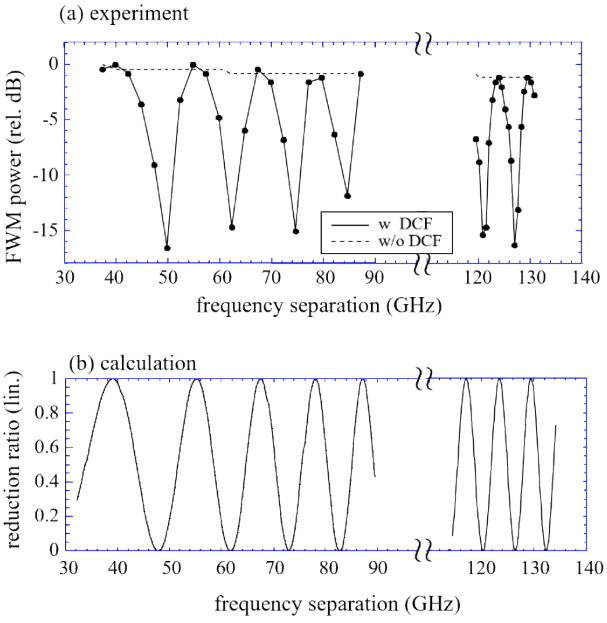


Fig. 8 FWM power as a function of frequency separation between signal and pump lights. DCF length was 0.5 km and signal light was CW.

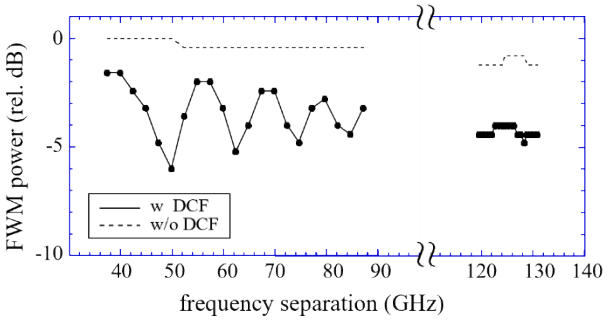


Fig. 9 FWM power as a function of frequency separation between signal and pump lights. DCF length was 0.5 km and signal light was QPSK modulated at 12.5 Gbps.

4. FWM Reduction in WDM System

The analytical formula indicating the FWM suppression is presented in Sect. 2, and its validity is experimentally confirmed in Sect. 3. Subsequently, this section calculates the FWM reduction ratio in an optically repeating WDM transmission system over DSFs, using the formula derived in Sect. 2 and experimentally confirmed in Sect. 3.

Based on Eq. (7), the total FWM power generated at the sth channel in a WDM transmission system with DEs can be expressed as

$$P_{\mathrm{F}} = 4\kappa \sum_{p,q,r} \eta\left(\Delta\beta_{pqr}\right) \frac{\sin^2\left[N\left(\Delta\beta_{pqr}L_0 + \Delta b_{pqr}L_{\mathrm{d}}\right)/2\right]}{\sin^2\left[\left(\Delta\beta_{pqr}L_0 + \Delta b_{pqr}L_{\mathrm{d}}\right)/2\right]} + \kappa \sum_{p,r} \eta\left(\Delta\beta_{ppr}\right) \frac{\sin^2\left[N\left(\Delta\beta_{ppr}L_0 + \Delta b_{ppr}L_{\mathrm{d}}\right)/2\right]}{\sin^2\left[\left(\Delta\beta_{ppr}L_0 + \Delta b_{ppr}L_{\mathrm{d}}\right)/2\right]}, \tag{21}$$

where $\{p, q, r\}$ indicate the channel numbers; $\kappa \equiv \gamma^2 P_0^3 \exp(-\alpha L_0)$; and the fiber dispersion is assumed to be uniform along the transmission lines, considering the worst-case scenario. The first summation represents the FWM power generated from nondegenerate processes satisfying $f_s = f_p + f_q - f_r$, and the second summation represents that from partially degenerate processes satisfying $f_s = 2f_p - f_r$.

Provided that DCFs of an appropriate length, with which the spectrum bandwidth of the FWM light is sufficiently wider than the frequency period, e.g., 2.0-km DCFs for 12.5 Gbaud QPSK systems, are used, Eq. (21) is averaged as follows:

$$\begin{aligned}\langle P_F \rangle &= 4\kappa \sum_{p,q,r} \eta\left(\Delta\beta_{pqr}\right) \left\langle \frac{\sin^2\left[N\left(\Delta\beta_{pqr}L_0 + \Delta b_{pqr}L_\mathrm{d}\right)/2\right]}{\sin^2\left[\left(\Delta\beta_{pqr}L_0 + \Delta b_{pqr}L_\mathrm{d}\right)/2\right]} \right\rangle \\ &\quad + \kappa \sum_{p,r} \eta\left(\Delta\beta_{ppr}\right) \left\langle \frac{\sin^2\left[N\left(\Delta\beta_{ppr}L_0 + \Delta b_{ppr}L_\mathrm{d}\right)/2\right]}{\sin^2\left[\left(\Delta\beta_{ppr}L_0 + \Delta b_{ppr}L_\mathrm{d}\right)/2\right]} \right\rangle \\ &= \kappa N \left\{ 4 \sum_{p,q,r} \eta\left(\Delta\beta_{pqr}\right) + \sum_{p,r} \eta\left(\Delta\beta_{ppr}\right) \right\}. \end{aligned} \tag{22}$$

Subsequently, the FWM reduction ratio for QPSK-modulated signal lights, $\langle R \rangle = \langle P_\mathrm{F}(L_\mathrm{d} \neq 0) \rangle / P_\mathrm{F}(L_\mathrm{d} = 0)$, is evaluated as

$$\begin{aligned}\langle R \rangle &= N \left\{ 4 \sum_{p,q,r} \eta\left(\Delta\beta_{pqr}\right) + \sum_{p,r} \eta\left(\Delta\beta_{ppr}\right) \right\} \\ &\quad \times \left\{ 4 \sum_{p,q,r} \eta\left(\Delta\beta_{pqr}\right) \frac{\sin^2\left[N\Delta\beta_{pqr}L_0/2\right]}{\sin^2\left[\Delta\beta_{pqr}L_0/2\right]} \right. \\ &\quad \left. + \sum_{p,r} \eta\left(\Delta\beta_{ppr}\right) \frac{\sin^2\left[N\Delta\beta_{ppr}L_0/2\right]}{\sin^2\left[\Delta\beta_{ppr}L_0/2\right]} \right\}^{-1}. \end{aligned} \tag{23}$$

The phase mismatch in this equation is expressed as [4]

$$\Delta\beta_{pqr} = \frac{\pi\lambda^4 D_\mathrm{cc}}{c^2}\left(f_p + f_q - 2f_0\right)(p - r)(q - r)\Delta f^2, \tag{24a}$$

$$\Delta\beta_{ppr} = \frac{2\pi\lambda^4 D_\mathrm{cc}}{c^2}\left(f_p - f_0\right)(p - r)^2\Delta f^2, \tag{24b}$$

where $D_\mathrm{cc} = dD_\mathrm{c}/d\lambda$ with D_c being the dispersion parameter of the DSFs; Δf is the channel frequency spacing; and f_0 is the zero-dispersion frequency of the DSFs.

Using Eq. (23), we calculated the FWM reduction ratio as a function of the number of spans for the center channel in a 100-GHz spaced 11-channel WDM system. The span length was 80 km, and the center channel was assumed to be positioned at the zero-dispersion wavelength of the transmission fibers, considering the worst-case scenario. The signal lights were assumed to have a single and identical polarization state. However, the result obtained under this condition of the polarization state would be applicable to polarization-multiplexed systems, because the FWM reduction mechanism of randomizing the relative phases of FWM lights generated in each span works independently on the two orthogonal polarization components and the FWM reduction is obtained in each polarization state.

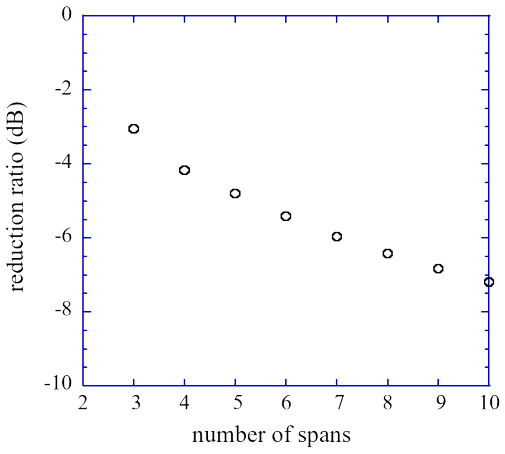


Fig. 10 FWM reduction ratio at center channel, positioned at zero-dispersion wavelength of transmission fibers, in 11-channel WDM system. Channel spacing was 100 GHz and repeater spans was 80 km.

The results are presented in Fig. 10. As shown, the FWM reduced efficiently as the number of spans increased. Although the previous section indicated that the reduction ratio was $1/N$ for a phase-matched FWM component in an N-span system, the reduction ratio shown in Fig. 10 did not reach this level. This is because FWM in WDM systems includes components that do not satisfy the phase-matching condition, i.e., $\Delta\beta_{pqr} \neq 0$, and the efficiency of FWM reduction for those components is not comparable to that for phase-matched components.

5. Conclusion

A scheme for FWM reduction in optically repeating WDM transmission systems over DSFs was presented. DEs such as DCFs were inserted at repeating points, through which the relative phase between the transmitted signal lights and FWM lights generated in the previous spans was shifted. Consequently, FWM lights generated in each span overlapped in random phases, and the total FWM power at the receiver was lower than that in systems with no DEs. Proof-of-principle experiments were conducted, and the results confirmed the FWM reduction mechanism above. Calculation for evaluating the reduction ratio in WDM systems was presented.

References

[1] G.P. Agrawal, Nonlinear Fiber Optics, Academic Press, San Diego, 2001.
[2] R.W. Tkach, A.R. Chraplyvy, F. Forghieri, A.H. Gnauck, and R.M. Derosier, "Four-photon mixing and high-speed WDM systems," J. Lightw. Technol., vol.13, no.5, pp.841–849, May 1995.
[3] K. Inoue, K. Nakanishi, K. Oda, and H. Toba, "Crosstalk and power penalty due to fiber four-wave mixing in multichannel transmissions," J. Lightw. Technol., vol.12, no.8, pp.1423–1439, Aug. 1994.
[4] K. Inoue, "Four-wave mixing in an optical fiber in the zero-dispersion region," J. Lightw. Technol., vol.10, no.11, pp.1553–1561, Nov. 1992.

[5] M. Jinno, T. Sakamoto, J. Kani, S. Aisawa, K. Oda, M. Fukui, H. Ono, and K. Oguchi, "First demonstration of 1580 nm wavelength band WDM transmission for doubling usable bandwidth and suppressing FWM in DSF," Electron. Lett., vol.33, no.10, pp.882–883, May 1997.
[6] K.O. Hill, D.C. Johnson, B.S. Kawasaki, and R.I. MacDonald, "cw three-wave mixing in single-mode optical fibers," J. Appl. Phys., vol.49, no.10, pp.5098–5106, Oct. 1978.
[7] K. Inoue and H. Toba, "Fiber four-wave mixing in multi-amplifier systems with nonuniform chromatics dispersion," J. Lightw. Technol., vol.13, no.1, pp.88–93, Jan. 1995.
[8] K. Inoue, "Phase-mismatching characteristic of four-wave mixing in fiber lines with multistage optical amplifiers," Opt. Lett., vol.17, no.11, pp.801–803, June 1992.
[9] N. Shibata, R.P. Braun, and R.G. Waarts, "Phase-mismatch dependence of efficiency of wave generation through four-wave mixing in a single-mode optical fiber," IEEE J. Quantum Electron., vol.QE-23, no.7, pp.1205–1210, July 1987.

Appendix:

In this section, we derive Eq. (16). First, we introduce variable $x = (\Delta\beta L_0 + \Delta b L_d)/2$ to simplify the left-hand side of Eq. (16) as follows:

$$S_N = \left\langle \frac{\sin^2\left[N\left(\Delta\beta L_0 + \Delta b L_d\right)/2\right]}{\sin^2\left[\left(\Delta\beta L_0 + \Delta b L_d\right)/2\right]} \right\rangle = \left\langle \left\{ \frac{\sin(Nx)}{\sin x} \right\}^2 \right\rangle. \quad (A\cdot 1)$$

For $N = 2$ and 3, S_N can be calculated as

$$S_2 = \left\langle \left\{ \frac{\sin(2x)}{\sin x} \right\}^2 \right\rangle = 4\left\langle \cos^2 x \right\rangle = 2 \quad (A\cdot 2)$$

and

$$S_3 = \left\langle \left\{ \frac{\sin(3x)}{\sin x} \right\}^2 \right\rangle = \left\langle \{1 + 2\cos(2x)\}^2 \right\rangle = 3. \quad (A\cdot 3)$$

Next, we calculate S_N for larger values of N, through which the following expressions are supposed to be satisfied:

$$\frac{\sin(Nx)}{\sin x} = 2\sum_{k=1}^{N/2} \cos[(2k-1)x] \quad (A\cdot 4a)$$

for even N, and

$$\frac{\sin(Nx)}{\sin x} = 1 + 2\sum_{k=1}^{(N-1)/2} \cos(2kx) \quad (A\cdot 4b)$$

for odd N.

Subsequently, we prove Eq. (A·4) using mathematical induction. First, Eq. (A·4) is assumed to be satisfied for $N-1$. Subsequently, $\sin(Nx)/\sin(x)$ is developed as follows:

$$\begin{aligned}\frac{\sin(Nx)}{\sin x} &= \frac{1}{\sin x}\{\sin[(N-1)x]\cos x + \cos[(N-1)x]\sin x\} \\ &= \left\{1 + 2\sum_{k=1}^{(N-2)/2} \cos(2kx)\right\}\cos x + \cos[(N-1)x] \\ &= \cos x + \sum_{k=1}^{N/2-1} \{\cos[2(k+1)x] + \cos[2(k-1)x]\}x \\ &\quad + \cos[(N-1)x] \\ &= 2\sum_{k=1}^{N/2} \cos[(2k-1)x] \end{aligned} \quad (A\cdot 5a)$$

for even N, and

$$\begin{aligned}\frac{\sin(Nx)}{\sin x} &= \frac{1}{\sin x}\{\sin[(N-1)x]\cos x + \cos[(N-1)x]\sin x\} \\ &= 2\sum_{k=1}^{(N-1)/2} \cos[(2k-1)x]\cos x + \cos[(N-1)x] \\ &= \sum_{k=1}^{(N-1)/2} \{\cos(2kx) + \cos[2(k-1)x]\} \\ &\quad + \cos[(N-1)x] \\ &= \sum_{k=1}^{(N-1)/2} \cos(2kx) + 1 + \sum_{k=2}^{(N-1)/2} \cos[2(k-1)x] \\ &\quad + \cos\left[2\cdot\frac{N-1}{2}\cdot x\right] \\ &= 1 + 2\sum_{k=1}^{(N-1)/2} \cos(2kx) \end{aligned} \quad (A\cdot 5b)$$

for odd N, where Eq. (A·4) with $N-1$ is applied. Equation (A·5) indicates that Eq. (A·4) is satisfied provided that it is satisfied for $N-1$. Futhermore, the validity of Eq. (A·4) for $N = 1$ and 2 are confirmed in Eqs. (A·2) and (A·3), respectively. Therefore, Eq. (A·4) is satisfied for any N.

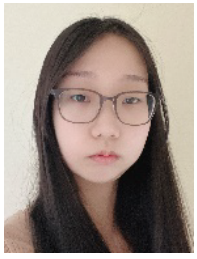
Using Eq. (A·4), we can derive Eq. (A·1) as

$$\begin{aligned} S_N &= 4\left\langle \left\{ \sum_{k=1}^{N/2} \cos[(2k-1)x] \right\}^2 \right\rangle \\ &= 4\left\langle \sum_{k=1}^{N/2} \cos^2[(2k-1)x] \right\rangle \\ &\quad + 8\left\langle \sum_{k<k'} \cos[(2k-1)x]\cos[(2k'-1)x] \right\rangle \\ &= N \end{aligned} \quad (A\cdot 6a)$$

for even N, and

$$\begin{aligned} S_N &= \left\langle \left\{ 1 + 2\sum_{k=1}^{(N-1)/2} \cos(2kx) \right\}^2 \right\rangle \\ &= \left\langle 1 + 4\sum_{k=1}^{(N-1)/2} \cos^2(2kx) \right\rangle \\ &\quad + 4\left\langle \sum_{k<k'} \cos(2kx)\cos(2k'x) \right\rangle \\ &= N \end{aligned} \quad (A\cdot 6b)$$

for odd N. Therefore, Eq. (16) is derived.



Ayano Inoue is an undergraduate student of the School of Engineering, Osaka University.

Koji Igarashi received the B.E. degree in electrical and computer engineering from Yokohama National University, Yokohama, Japan, in 1997, and the M.E. and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1999 and 2002, respectively. He was with the Furukawa Electric Corporation, Ltd. from 2002 to 2004, the University of Tokyo from 2004 to 2011, and KDDI R&D Laboratories, Inc. from 2012 to 2013. He is currently a Professor at Osaka University, Osaka, Japan. His current works include high-capacity long-haul optical fiber transmission systems, signal processing for coherent optical communication systems, and optical fibers devices for space-division multiplexed optical transmission systems.

Shigehiro Takasaka received B.S., M.S. and the Ph.D. degrees in Physics from Hokkaido University, Sapporo, Japan, in 1994, 1996, and 1999, respectively. He was a postdoctoral fellow of The University of Tokyo from 1999 to 2002. He joined Furukawa Electric Co., Ltd., Tokyo, Japan in 2002. He has been engaged in research and development of optical fiber applications such as nonlinear optical signal processing and fiber amplifiers. He was also a manager of Sakigake project on PRESTO, JST, Kawaguchi, Japan from 2004 to 2008. He is a senior researcher of optical application section of Photonics Laboratories., and a senior member of IEEE photonics society.

Kyo Inoue received the B.S. and M.S. degrees in applied physics in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from Tokyo University, Tokyo, Japan, in 1997. From 1984 to 2005, he was with Nippon Telegram and Telephone Corporation, where his work involved optical communications and quantum communications. He is currently a Professor at Osaka University, Osaka, Japan.

PAPER

Flexi COCOA: Flexible Weight Based RTO Computation Approach for COCOA

Archana K. RAJAN[†a)], ***Nonmember and*** **Masaki BANDAI**[†b)], ***Member***

SUMMARY Constrained Application Protocol (CoAP) is a popular UDP based data transfer protocol designed for constrained nodes in lossy networks. Congestion control method is essentially required in such environment to ensure proper data transfer. CoAP offers a default congestion control technique based on binary exponential backoff (BEB) where the default method calculates retransmission time out (RTO) irrespective of Round Trip Time (RTT). CoAP simple congestion control/advanced (COCOA) is a standard alternative algorithm. COCOA computes RTO using exponential weighted moving average (EWMA) based on type of RTT; strong RTT or weak RTT. The constant weight values used in COCOA delays the variability of RTO. This delay in converging the RTO can cause QoS problems in many IoT applications. In this paper, we propose a new method called Flexi COCOA to accomplish a flexible weight based RTO computation for strong RTT samples. We show that Flexi COCOA is more network sensitive than COCOA. Specifically, the variable weight focuses on deriving better RTO and utilizing the resources more. Flexi COCOA is implemented and validated against COCOA using the Cooja simulator in Contiki OS. We carried out extensive simulations using different topologies, packet sending rates and packet error rates. Our results show that Flexi COCOA outshines COCOA and can improve QoS of IoT monitoring applications.

key words: *IoT, application protocol, CoAP, COCOA, congestion control*

1. Introduction

In 1999, the term "Internet of Things", abbreviated as IoT, was introduced to the world for the first time by the British computer scientist Kevin Ashton [1]. Ever since then, IoT has been a buzzword which is driving academic and industrial attention alike. Now, with the huge number and types of devices connected to the IoT, it has become an inevitable part of all walks of life [2]. Devices connected in the IoT, referred as nodes, are constrained in memory, processing, and power. Also, they operate in lossy networks with narrow bandwidth. These special characteristics can often cause low throughput and large packet errors. IETF has been focusing on deriving protocols suitable for such environments. Their contributions include constrained application protocol (CoAP) [3], routing protocol for low power lossy network (RPL) [4], IPv6 over low power wireless personal area network (6LoWPAN) [5].

Congestion is an undesired event when the node or link propagates more packets than it can handle. Considering the constrained capacities of nodes and networks, congestion is a critical problem in IoT. In general, constrained devices in IoT networks exchange small amounts of data periodically. When many such devices interact simultaneously, demand of resources exceeds its availability and results in congestion. Consequences of congestion are packet drop from buffer, increased queuing delay, decreased throughput etc. [6]. Therefore, congestion control (CC) technique is necessarily deployed to detect and reduce congestion.

CoAP is one of the promising UDP based application layer protocols for IoT. When reliability is required by the application, CoAP includes confirmable message and acknowledgement. When the ACK is not received at the application layer in stipulated time, the CC method is triggered. Therefore, the waiting time for the ACK, referred to as retransmission time out (RTO), needs to be known at the application layer [7].

A static value of RTO is insufficient for catering to the needs of the dynamics of the network. Instead, deriving an RTO, aligned with the congestion level of the network, ought to be introduced in CoAP [8]. The attributes which signal the congestion level need to be introduced to make the RTO more responsive and accurate. However, tracking, computation and storing of attributes that signal congestion shall not overhead the constrained nodes.

The default CC method in CoAP computes RTO using binary exponential backoff (BEB). The simplicity of this method also has a downside that RTO is insensitive to congestion in the network. Such insensitivity can result in small RTOs at high congestion levels and large RTOs at low congestion levels [9]. This in turn makes the node inappropriately aggressive and conservative. Hence the default CC method is inefficient to deliver data promptly to the server node. As a consequence, the efficiency of many crucial monitoring applications in health care, disaster management, pollution tracking etc. can be adversely impacted.

CoAP simple congestion control/advanced (COCOA) is the most popular alternative proposed [9] and later revised [10]. Round trip time (RTT) is used to compute RTO by means of exponential weighted moving average method (EWMA). RTT samples computed from packets delivered with and without retransmissions result in two types of RTT: namely weak RTT and strong RTT. Each of them is used to derive weak RTO and strong RTO respectively. The overall RTO estimate of COCOA is updated when either weak RTO or strong RTO is updated.

Both RTT and its type are indicators of congestion level. Increase in RTT indicates an increase in congestion and vice

Manuscript received February 1, 2024.
Manuscript revised May 15, 2024.
Manuscript publicized June 28, 2024.

[†]Faculty of Science and Technology, Sophia University, Tokyo, 102-8554 Japan.
a) E-mail: archanakrajan@gmail.com
b) E-mail: bandai@sophia.ac.jp

DOI: 10.23919/transcom.2024EBP3023

versa. Similarly, the type of RTT samples shows if sender runs into retransmission, which implies whether congestion being built up in the network. Therefore, both RTT and type of RTT help to derive RTO sensitive to congestion. Many evaluations show that COCOA outperforms the default CC method in CoAP [9], [10].

However, on the other hand, the constant weight values in the EWMA based RTO estimation create concern about the performance. The constant weight values always make the current sample contribute the same share irrespective of changes in the congestion. This causes a delay in fluctuating RTO sensitively. Weight values in the RTO estimations can also contribute to the effectiveness of COCOA if they are network sensitive. Therefore, the idea of adaptive weight value setting in RTO estimation should be examined.

In this paper, we propose a novel approach called Flexi COCOA, for computing RTOs using adaptive weight value for strong RTO. We show that the proposed method is more effective than COCOA to derive better RTO, more packet transmissions, and less retransmissions for periodic data transfer nodes. The main contributions of this paper are as follows.

- We introduce, to our knowledge, the first method of considering ratio of RTTs as measure for the weight value estimation of strong RTO.
- The proposed method is implemented in Contiki Operating System with minimal changes in COCOA algorithm.
- We demonstrate the effectiveness of the proposed method on Cooja simulator.

The rest of the paper is organized as follows. We describe the details of CoAP in Sect. 2 and related work in Sect. 3. The proposed method is presented in Sect. 4. Details about evaluation and results are presented in Sect. 5. We conclude this paper with future research directions in Sect. 6.

2. CoAP Protocol

CoAP [3] is designed to be a lightweight application layer protocol adhering to request-response model following representational state transfer (REST) architecture. Requests are originated by the client and destined to the server. On receiving a request, server originates responses that are destined to the client. Therefore, for the CoAP messaging model, in the context of requests, client and server are sender and receiver respectively. Similarly, for response, server and client are sender and receiver respectively. Like HTTP, CoAP uses the methods such as GET, POST, PUT and DELETE to access or manipulate resources on the server. Four message types are defined by this protocol; namely non-confirmable (NON), confirmable (CON), reset (RST) and acknowledgment (ACK).

NON request message is used when reliability is not expected via acknowledgments. CON request messages are always coupled with ACK messages from the receiver and hence provide reliability. Message id helps to match CON message and ACK message. Sender can attempt, by default, up to four retransmission attempts if ACK is not received in stipulated time. This waiting time is derived by an RTO estimator of the CC method. A conservative restriction is set on sender to have only one request per destination to have no ACK received yet. When the message is received in error, the receiver replies with RST.

RFC 7252 [3] suggests a trivial default CC method of CoAP. A random RTO is chosen from [2, 3] seconds for every new CON transmission. A timer starts at the sender to follow RTO. If the timer expires and ACK is not received, then a packet loss is assumed. The sender attempts retransmission of the same packet, after BEB is applied to RTO. BEB is applied on RTO to double its value for each retransmission as shown in Eq. (1).

$$RTO_{\text{current}} = RTO_{\text{previous}} \times 2. \tag{1}$$

The advantage of default CC is that there is no need to manage information about the end-to-end data transfer. However, the simplicity of the default method also results in a serious drawback. Computation of RTO does not keep track of congestion in the network. Therefore, an estimated RTO may become an overestimate or underestimate. In case of an underestimated value, the sender spuriously transmits packets into the constrained network, worsening the congestion. An overestimated value makes the sender transmit less packets and underutilize resources. Consequently, a network sensitive algorithm is necessary.

3. Related Work

COCOA is a network sensitive end-to-end CC technique for CoAP message transmission. The first version of COCOA is presented in [9]. COCOA+ [10] is the latest version with minor improvements on [9]. Henceforth, we use the word COCOA instead of COCOA+. In this section, we present the working of COCOA algorithm and different enhancements on COCOA.

3.1 Congestion Control in COCOA

COCOA [10] is an improvement over the default CC method of CoAP by associating RTT value with RTO computation. It follows the principles of the Karn-Patridge algorithm [11] as in TCP, to estimate RTO from RTT. TCP assumes packet loss is caused by network congestion and for this reason ignores RTT samples of retransmitted packets. However, packet loss is expected in lossy network by high bit error rate (BER) also. Therefore, COCOA pays attention to retransmitted packets.

RTT sample of the packet which is acknowledged without retransmission is called strong RTT and that with retransmission is called weak RTT. Strong RTO is computed from strong RTT, and weak RTO is computed from weak RTT. A precise mapping of whether the ACK corresponds to original CON message or retransmitted copy, is not possible. Therefore, the number of retransmissions in weak RTT

measurement is limited to two. Further retransmissions, if any, are avoided in estimation to reduce the precision error.

For each CoAP receiver, a CoAP sender needs to maintain two types of state variables for all the related measures as well. When a strong or weak RTT is measured, the corresponding estimator gets updated. Let rtt and x denote the RTT value measured and the type of RTT respectively. Smoothed RTT, smoothed variance of $RTT(RTVAR)$, and RTO are updated as follows.

$$RTT_x = (1-\alpha) \times RTT_x + \alpha \times rtt \quad (2)$$

$$RTVAR_x = (1-\beta) \times RTVAR_x + \beta \times |RTT_x - rtt| \quad (3)$$

$$RTO_x = RTT_x + K_x \times RTVAR_x \quad (4)$$

COCOA follows the recommendation of RFC 6298 [12] to assign the value of α, β and K_{strong} as 0.125, 0.25 and 4 respectively. The value of K_{weak} is assigned to be 1 to reduce the inaccuracy from ambiguity issues. New RTO is calculated as given below, with $weight_{strong}$ as 0.5 and for $weight_{weak}$ as 0.25.

$$RTO_{new} = weight_x \times RTO_x + (1 - weight_x) \times RTO_{new} \quad (5)$$

Using a dithering technique, initial RTO for the next transmission, RTO_{init}, is selected from RTO_{new}.

$$RTO_{init} = [RTO_{new}, 1.5 \times RTO_{new}] \quad (6)$$

In case of retransmissions, COCOA considers a variable backoff mechanism (VBF) to calculate new RTO value, $RTO_{current}$.

$$RTO_{current} = RTO_{previous} \times VBF \quad (7)$$

This is to control the RTO value becoming smaller or larger. VBF is 3 if RTO_{init} is less than 1 second, and it is 1.3 when RTO_{init} greater than 3 seconds. Otherwise VBF is 2, which is normal BEB.

COCOA also introduces a mechanism to age the estimations in case RTT samples are not available for a certain duration. Motivation is that, in a lossy environment, it is normal not to get samples for a long period and the past values become stale. Therefore, the aging mechanism is required to converge the estimation into default values.

In general, COCOA incorporates congestion level indicators such as RTT, variation in RTT and retransmissions for RTO computation. Therefore, COCOA is advantageous over the default CC method, with the ability to adjust RTO dynamically to network changes. Evaluations in [9], [10] and [13] compare COCOA against default CoAP.

3.2 Varaints of COCOA

Multiple solutions are available to address different deficiencies of COCOA [14]–[29]. We summarize them in Table 1. Most of the methods in Table 1 depend on constant weight values for RTO estimation. COCOA suffers from a critical drawback due to the constant weight values used in the RTO computation. Irrespective of the network dynamics, constant weight values force the current RTT sample to always contribute identically. This can slow down the fluctuation of RTO and thereby affect the data delivery of periodic monitoring applications. Thus, a network sensitive weight value is necessary to accelerate the RTO fluctuation. Our analysis about existing solutions addressing the constant weight problem is given below.

AdCOCOA [21] suggests all parameters being dynamic. Weak RTO estimator is eliminated in AdCOCOA; however weak RTT samples are considered. The difference between smoothed RTT and current RTT is taken as an indication of network status and is used to derive the multipliers for scaling the parameters. Results show that lower RTO and higher packet sending rate with lesser retransmissions are achieved. When the RTT is measured after retransmissions, the scaling factors can grow larger at once. Therefore, a larger RTO results which further can cause long idle period. Bounding of scaling factor is important in this aspect; however it is not included. CoAP recommends keeping message overhead small, thus limiting the need for fragmentation at lower layers. Fragmentation reduces the probability of packet delivery. Loss or delay of even a single fragment can eventually cause the original CON packet to be resent. In the constrained environment, such retransmissions induce additional traffic that may worsen congestion level of the network. It is noteworthy that layer 2 packet size is limited to 127 bytes by 6LoWPAN [5]. A payload must fit into a single 6LoWPAN frame to avoid fragmentation. Simulation parameters in AdCOCOA shows that the packet size used in the experiments is 1010 bytes. Such a packet size can cause fragmentation problems at 6LoWPAN. Therefore, block-wise trasnsfer method would be opted for larger packet size.

mlCOCOA [24] estimates parameters using support vector machine (SVM) based on client number, packet size and packet delivery ratio. The predicted values are used in the COCOA in place of their default values. mlCOCOA achieves higher throughput in comparison to CoAP and COCOA. A major drawback of mlCOCOA is that it demands clients with plentiful resources and hence not suitable for constrained devices. Prediction based on client number also raises questions. In case an expected node does not participate in RPL due to error, or power outage or mobility, the predicted values may not be suitable for the network. Also, the learning phase from the training set is static. This can be different from dynamic evaluations.

GSRTC [26] enhances COCOA in lossless network conditions. GSRTC derives a geometric series based weight for strong RTO. Weight value is updated on each consecutive strong RTT. When a strong RTT sample is obtained, $weight_{strong}$ is derived from Eq. (8).

$$weight_{new} = weight_{prev} + (\frac{1}{2})^{count} \quad (8)$$

$weight_{new}$ is used as $weight_{strong}$ in Eq. (5),

Table 1 Summary of COCOA variants.

Algorithm	Defect Identified in COCOA	Solution	Weight Value
CoCoA4-state-Strong [14]	Uniformly considering all retransmission attempts	RTO computation is separated for each retransmission attempts	Fixed
COCOA-AI [15]	Threshold of Variable Back Off and aging	Threshold is reset with respect to RTOstrong, Message id is changed for retransmission	Fixed
FASOR [16]	Dealing link-error packet loss	Different RTO estimators for different states of congestion	Fixed
Full Backoff [17]	Addressing bufferbloated bottleneck buffer	Retain the back-off RTO values	Fixed
pCOCOA [18]	Inaccuracy of weak estimator and RTTVAR	Precise match of CON- ACK, Maximum deviation method for RTTVAR	Fixed
COCOA++ [19]	Per packet RTT analysis	CAIA Delay Gradient	None
COCOA-RED [20]	Packet loss of group communication of OBSERVE	Use Random Early detection algorithm and a Fibonacci series based Pre-Increment Backoff (FPB)	None
AdCOCOA [21]	Constant weight values	Removed weak estimator, Variable weights are derived using change of RTT values	Variable
CACC [22]	Differentiate the loss of packets	Use three types of RTO estimators	Fixed
Genetic COCOA++ [23]	Per packet RTT analysis	Delay gradient along with the genetic algorithm and probabilistic backoff factor	Fixed
mlCOCOA [24]	Constant weight values	Variable weight values are derived using machine learning technique	Variable
RCOAP [25]	Support to burst data traffic	Transmission rate is adjusted according to band width delay product	Fixed
GSRTC [26]	Constant weight values	Retain back-off RTO values, Variable weights are derived using consecutive strong RTTs	Variable
psoCOCOA [27]	Default value for maximum number of retransmissions and age of RTO value	Particle swarm optimization (PSO) is used to tune the parameters	Fixed
6COCOA [28]	Wireless communication in 6TiSCH	Parameter values are reset and uses MSF	Fixed
iCOCOA [29]	Reliability on RTT	Use deep reinforcement learning and considers more features than COCOA	Fixed

$weight_{prev}$ denotes the weight obtained from the previous strong RTT sample and *count* indicates count of consecutive strong RTT samples. When a weak RTT sample is acquired, GSRTC adopts the *Fullbackoff2* variant of COCOA [17]. With *Fullbackoff2*, GSRTC selects RTO for the subsequent transmission as maximum among backed off RTO and RTO_{new}. Furthermore, $weight_{prev}$ and *count* are reset to 0 resulting in an aggressive reduction of $weight_{strong}$.

GSRTC achieves better flow completion time, better throughput, and lower number of retransmissions in the performance evaluation. Consecutive strong RTT samples are mandatory for GSRTC to improve over COCOA. A constrained network cannot guarantee consecutive strong RTT samples. Aggressive increase of weight value is exhibited by GSRTC whenever it receives consecutive strong RTT samples. As a result, RTO value becomes smaller quickly and cause a larger number of packets to reach the network in a short span of time. This can cause overhead in constrained nodes. Another drawback is the quick reduction of $weight_{strong}$ when a weak RTT is received. Since packet collisions can also result in weak RTT, a quick reduction of weight value may overestimate the congestion level of the network. Such an approach can degrade the rate of RTO fluctuation.

In general, existing solutions addressing the constant weight value problem achieve better performance than

COCOA. However, each of such solutions compromise the general characteristics of data transfer for constrained nodes outlined in [3], such as small payload size, resource bounded devices or lossy network conditions. Therefore, a new solution, adhering to the general characteristics, needs to be developed.

4. Proposed Method

In this paper, we propose Flexi COCOA through its motivation, overview, algorithm, and an illustration of a sample client-server interaction.

4.1 Motivation

In [9], the authors indicate about improving the performance of COCOA using adaptive parameter setting having tradeoff with complexity. COCOA algorithm makes use of many constant parameter values such as α, β, K_{strong}, K_{weak}, $weight_{strong}$ and $weight_{weak}$. However, the contribution of COCOA is confined only to K_{weak}, $weight_{weak}$ and $weight_{strong}$. The rest of the values are inherited from TCP. We aim at improving the parameters contributed by COCOA. The accuracy of weak RTT is a hindrance for making K_{weak} and $weight_{weak}$ to be adaptive. Modification of K_{weak} and $weight_{weak}$ is improper without addressing the precision of the weak RTT sample. Our analysis, thus, focuses on improving the weight associated with strong RTT sample, $weight_{strong}$.

4.2 Overview

Flexi COCOA enhances the RTO computation of COCOA by flexible weight derived using strong RTT samples. Gradual adjustment of weight is favored in Flexi COCOA to prevent overreactions such as those seen in GSRTC. Upon receiving a strong RTT sample, Flexi COCOA analyzes the trend in their count and revises the weight value accordingly. The revised weight aims to reduce transmission delay proportional to the number of packet losses.

The core idea is to increase and decrease the weight value of RTO_{strong} in Eq. (5) gradually based on the count of strong RTT samples to achieve faster RTO adaptation than COCOA. An increase in the count of strong RTT samples always signals less congestion. Therefore, weight value should be increased to provide more weightage to RTO_{strong} and produce small RTO for subsequent transmission. A gradual increase is preferable to prevent spurious transmissions. A decrease in the count of strong RTT samples can be caused by congestion, bit errors, or both. Flexi COCOA, like other COCOA variants, doesn't engage with lower layers of the protocol stack to identify packet loss causes, to preserve the lightweight application layer. While a reduction in weight value is necessary, resetting the weight value to a pre-determined constant value is improper as the cause of loss is unidentified. Therefore, a gradual weight reduction is preferred for RTO_{strong} to prevent overestimation of RTO.

A network sensitive weight value thus requires incorporating strong RTT samples observed, to incorporate fluctuation of congestion, and evolving cautiously. Therefore, we define weight value as the ratio between count of strong RTT samples and count of total RTT samples. The proposed weight value speeds up RTO convergence to the actual RTT compared to COCOA. This modification addresses transmission delay relative to the number of packets experiencing timeout retransmission. The advantage of such an RTO convergence is vital when strong RTT samples are collected after a weak RTT sets RTO to be large value. At any point in time, the proposed weight value is dynamic and considers the overall performance of client-server message exchange rather than that of a brief time.

Flexi COCOA makes minimal changes to the standard COCOA. Hence, resourceful devices are not a necessary requirement for the implementation of Flexi COCOA. Also, any stringent condition such as consecutive strong RTT samples is not mandated by Flexi COCOA to deliver desired performance.

4.3 Algorithm

Algorithm 1 shows the functioning of Flexi COCOA. Flexi COCOA is prevented from updating default COCOA weight, 0.5, for the first strong RTT sample for a cautious start-up of packet transmission. We introduce two new variables to the existing COCOA method: *count_RTT* and *count_SRTT*. *count_RTT* monitors aggregate count of RTT samples regardless of their type and *count_SRTT* exclusively monitors count of strong RTT samples. Both are initialized as zero.

count_RTT is updated for each RTT sample as shown in line 6 of the algorithm. When a weak RTT sample is collected, the proposed method inherits the weight value of 0.25 from COCOA. If the RTT sample is identified as strong RTT, *count_SRTT* is updated by the algorithm in line 12. For every strong RTT sample, the weight value is revised as the fraction of *count_SRTT* over *count_RTT* as in line

Algorithm 1 Flexi COCOA

1: **Initialize**
2: $weight \leftarrow 0.5$ ▷ fixed for first strong RTT
3: $count\_SRTT \leftarrow 0$
4: $count\_RTT \leftarrow 0$
5: **for all** $RTTcomputed$ **do**
6: $count\_RTT++$ ▷ count of RTT is increased
7: **if** ACK received after retransmission **then**
8: $type \leftarrow weak$
9: $weight \leftarrow 0.25$
10: **else**
11: $type \leftarrow strong$
12: $count\_SRTT++$ ▷ count of strong RTT is increased
13: $weight \leftarrow \frac{count\_SRTT}{count\_RTT}$ ▷ weight is revised
14: **end if**
15: $RTT_{type} \leftarrow (1-\alpha) \times RTT_{type} + \alpha \times RTT$
16: $RTVAR_{type} \leftarrow (1-\beta) \times RTVAR_{type} + \beta \times |RTT_{type} - RTT|$
17: $RTO_{type} \leftarrow RTT_{type} + K_{type} \times RTVAR_{type}$
18: $RTO_{new} \leftarrow weight \times RTO_{type} + (1-weight) \times RTO_{new}$
19: **end for**

13 of the algorithm. We bound the value of this weight as, $0 < weight < 1$, at any instance. Lines 15 to 18 of the algorithm correspond identically to Eq. (2) through Eq. (5). The revised weight value provides flexible weightage to RTO_{strong}, while computing RTO_{new} as in line 18 of the algorithm. *count_RTT* and *count_SRTT* are reset to initialization if RTO value is reset to default initial value by invoking RTO aging technique. Flexi COCOA relies on *VBF* of COCOA to compute RTO value for retransmission as per Eq. (7).

COCOA, with constant weight, fluctuates the RTO at a steady rate. This can cause transmission delay. RTO adaptation using the network sensitive weight value of the proposed method can impact transmission delay in the following ways. Weight value in line 13 increases as the count of strong RTT sample increases. As a result, RTO_{strong} gains more weightage in line 18. RTO_{new}, derived with such higher weightage on RTO_{strong}, approaches to actual RTT faster than RTO_{new} from COCOA.

On the contrary, weak RTT sample normally sets theRTO_{new} to be large. Flexi COCOA benefits from adaptive weight value, when strong RTT samples are received after weak RTT sample. If weak RTT sample was a result of congestion, receiving strong RTT samples afterwards can be seen as an improvement in congestion. If weak RTT sample was caused by collision, receiving strong RTT samples afterwards shall address the overly setRTO_{new} by the weak RTT sample. In both cases, a faster convergence of RTO to RTT is the reaction required from the CC algorithm. Flexi COCOA exhibits faster RTO convergence compared to COCOA until it reaches a weight of 0.5, which is the weight used in COCOA. Weight value of Flexi COCOA remains larger than 0.5, until the count of weak RTT samples surpasses the count of strong RTT samples.

By offering accelerated RTO convergence, Flexi COCOA addresses transmission delay better than COCOA. As a result, clients can transmit a greater number of request packets in comparison to COCOA. During periods of severe congestion, the likelihood of obtaining a strong RTT sample is minimal, leading to a higher count of weak RTT samples compared to strong RTT samples. This can hinder the efficacy of Flexi COCOA. However, during periods of severe congestion, minimizing transmission delays may not be a significant concern. In such cases, the effectiveness of Flexi COCOA may not be a critical factor.

GSRTC [26] is a comparable solution to Flexi COCOA, as it adapts the weightage of RTO_{strong} based on strong RTT samples. GSRTC incorporates network dynamics exclusively from the lossless duration and therefore Eq. (8) aggressively boosts the weight of GSRTC to reduce RTO rapidly. Such an RTO reduction can lead to spurious transmissions. The weight updating process in Flexi COCOA progresses gradually, as instances of losses also impact the weight through denominator of the fraction in line 13 of Algorithm 1. This helps prevent spurious transmissions compared to GSRTC. Upon receiving a strong RTT sample after a weak RTT sample, GSRTC consistently resets the weight to 0.5. In such situations GSRTC is only capable of delivering RTO convergence rate equivalent to that of COCOA. In a comparable scenario, Flexi COCOA delivers superior performance than GSRTC. The weight value of Flexi COCOA may reach 0.5 only in congested periods where the speed of convergence is not a significant factor.

4.4 Illustration

We illustrate the weight updating using Flexi COCOA with the help of sample client-server interaction. Referring to Fig. 1, a client sends CON message to server and server sends back ACK message. The client receives ACK without retransmission for all packets except the second and sixth. Second and sixth packet attempted retransmissions and receive ACK. Therefore, RTT computed for second and sixth packets are weak RTT samples and others are strong RTT samples. Client updates the corresponding variables after receiving each ACK.

count_RTT is increased for all RTT samples and *count_SRTT* is updated for all strong RTT samples. The first strong RTT sample uses a weight of 0.5. Thereafter Flexi COCOA revises weight value. Flexi COCOA retains the default weight of COCOA for weak RTO computation. All updates of Flexi COCOA are highlighted in Fig. 1. Weight values are rounded to two decimal places for ease. Except for the first strong RTT, Flexi COCOA results in a larger weight value against 0.5 of COCOA. Also, it is evident that the occasional drops of second and sixth packets do not downgrade the $weight_{strong}$ to as low as 0.5 of COCOA.

5. Performance Evaluation

Flexi COCOA is implemented in Contiki-3.0 and simulated with Cooja simulator [30], [31]. Modifications are made to the er-cocoa.c file in the implementation of COCOA shared by the authors. We compare Flexi COCOA with COCOA and GSRTC. Apart from being the latest in existing methods, GSRTC can work on constrained nodes and retains the notion of strong RTT and weak RTT with clearly defined actions on each type of RTT. Also, GSRTC leverages the improvements already made in COCOA through *Fullbackoff*2. Therefore, we choose GSRTC from existing methods described under Sect. 3.2. This section provides details about the experiments set up, result analysis and discussion about the result.

5.1 Experiment Setup

Experiments are conducted on Cooja simulator which supports emulation of off-the shelf sensor node hardware by including the real specific node hardware. The compiled binary image file is to be uploaded into the simulated nodes, where the compiled code is then executed with the emulation model of the selected node type during simulations. At the physical and MAC layers, the nodes utilize IEEE 802.15.4, with a data transmission speed of 250 kbps in the 2.4 GHz radio frequency band.

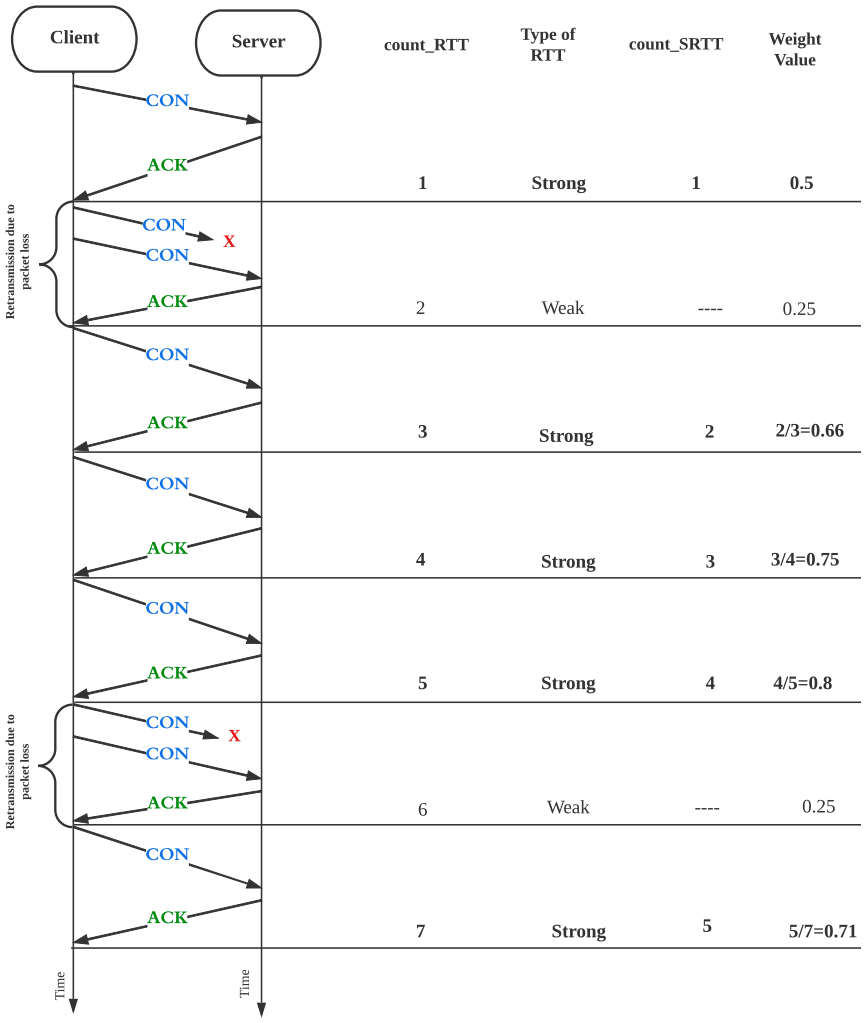
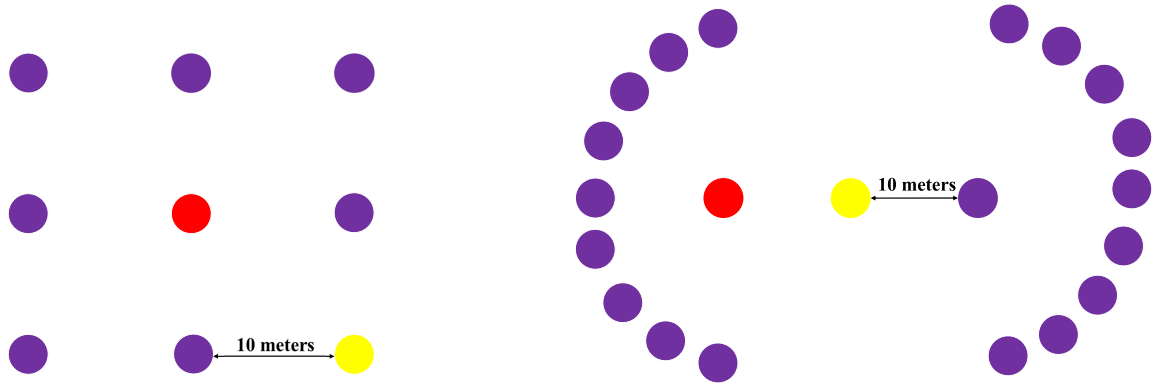


Fig. 1 Illustration of Flexi COCOA.



(a) Grid (b) Dumbbell

Fig. 2 Topologies.

Topologies used in the evaluation are Grid of 9 nodes and Dumbbell of 21 nodes as shown in Fig. 2. Grid and Dumbbell are widely used in evaluation of CC algorithms in CoAP [10], [14], [21]. In both topologies, one node is border router, one node is server, and the rest of the nodes are clients. Among the nodes in Fig. 2, the red color node is

Table 2 Sensor mote configuration.

Item	T mote Sky	Z1
RAM	10 kb	8 kb
ROM	48kb	98kb
MCU	MSP430F1611	MSP430F2617
Radio	CC 2420	

Table 3 Simulation parameters.

Parameter	Value
Operating System	Contiki 3.0
Simulator	Cooja
Radio Medium	Unit Disk Graph Medium Distance Loss
RDC Driver	NullRDC
MAC Driver	NullMAC
Mote Type	T-mote Sky, Zolerita Z1
Node Transmission Range	10 meters
Node Interference Range	20 meters
Topology	Grid of 9 nodes Dumbbell of 21 nodes
Packet Frequency (in milliseconds)	125, 250, 500, 1000
Packet Error Rate	0%,5%,10%
Duration of Simulation	7 minutes

the border router, the yellow color node is the server, and the rest of the nodes are clients. For the simulation, T-mote Sky mote [32] is used to configure border router and Zolerita Z1 mote [33] is used to configure rest of the nodes.

Table 2 shows the mote specification. Border router is responsible for initiating and collecting the RPL protocol messages and storing the routing information of all nodes. Therefore, the RAM requirement of border router is higher than other types of nodes in the network. Therefore, Sky mote is selected to configure border router. Z1 motes offer large ROM capacity which is beneficial for us to code applications and implement CC algorithm. Hence, Zolerita Z1 mote is selected to configure server and clients. Every client waits for 20 seconds, including the RPL to set up. Thereafter, clients send CoAP POST request to the server at specified frequency.

Simulation parameters of the experiments are summarized in Table 3. Cooja's Unit Disk Graph Medium (UDGM) with distance loss model is used for radio transmissions. The transmission range and the interference range of nodes are set to 10 meters and 20 meters respectively. When a node transmits packets, nodes in transmission range can receive those packets and nodes in interference range cannot. However, nodes in interference range can cause collisions if there are simultaneous packets. For UDGM, we can manually set the packet error rate (PER) in Cooja. Transmission ratio and reception ratio are provided in the user interface to introduce error. We use $0\%, 5\%, 10\%$ error rates to compare the performance of COCOA, GSRTC and Flexi COCOA.

Radio duty cycling handles the sleep and wakes up cycle depending on the type of the RDC method used by the motes. Motes are configured without Radio Duty Cycling (NullRDC), to never switch the radio off and to keep the motes always in listening mode. We use NullMAC as the MAC driver instead of traditional CSMA to prohibit retransmission mechanism at layer 2. MAC layer retransmissions, which is optional in IEEE 802.15.4, may improve the performance of a collision-prone network. By disabling MAC layer retransmissions, we simulate an increased risk of degraded network performance due to a higher likelihood of packet loss and subsequent retransmissions initiated from the application layer. The capability of CC algorithm to manage and mitigate packet losses in such a worst-case scenario can be thus evaluated.

Each topology is simulated using each packet transmission frequency and PER combination. The test is repeated 12 times with different random seeds. An average of each performance metric from them is presented in this paper.

5.2 Result Analysis

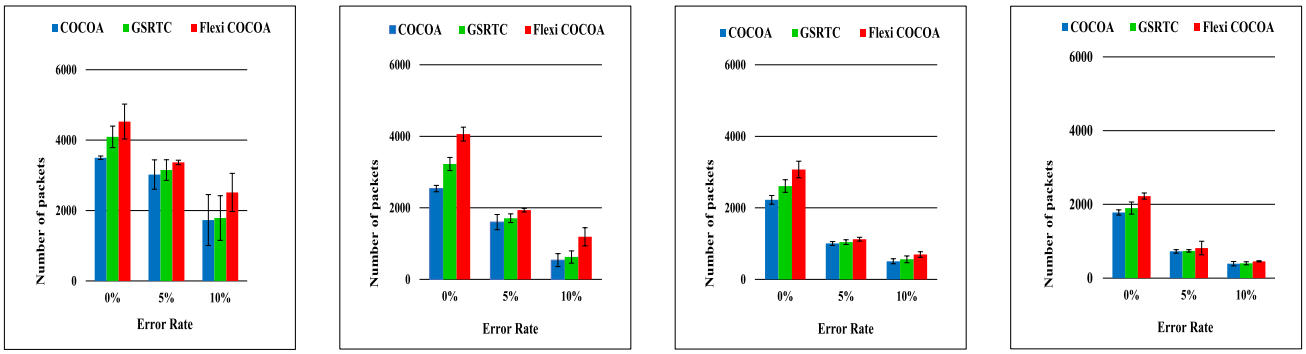
We compare the performance of Flexi COCOA with COCOA and GSRTC based on total packet transmissions and retransmission effort.

5.2.1 Total Packet Transmissions

We denote total packet transmissions as the number of packets successfully delivered from the clients to the server. Figures 3 and 4 report the total packet transmissions. We do not consider retransmitted copy of the packets in this count. Each result is presented with 95% confidence interval. In every scenario, Flexi COCOA achieves a significant increase in total packet transmissions than COCOA and GSRTC. Consequently, Flexi COCOA can significantly enhance the efficiency of periodic data monitoring applications by offering greater data transfer competence. Achievement in total packet transmission is correlated with addressing transmission delay. RTO value provides insight into managing transmission delay. Therefore, we analyze the RTO value obtained by each algorithm.

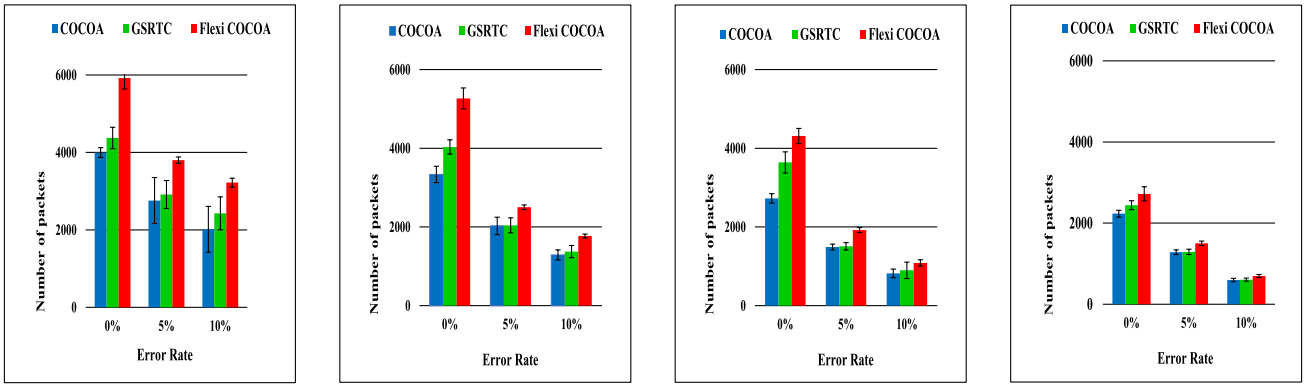
Tables 4 and 5 present the RTO values obtained from COCOA, GSRTC and Flexi COCOA. RTO values are in unit of seconds. Results in Tables 4 and 5 reveal that GSRTC and Flexi COCOA, with the help of variable weight parameter, derive lower RTOs than COCOA in all our experiments. Adaptive weight values of GSRTC and Flexi COCOA make the larger RTOs converge faster than COCOA. As a result, GSRTC and Flexi COCOA exhibit a larger total packet transmissions compared to COCOA.

The average values of adaptive weight from GSRTC



(a) Frequency: 125ms (b) Frequency: 250ms (c) Frequency: 500ms (d) Frequency: 1000ms

Fig. 3 Total Packet Transmissions of Grid Topology.



(a) Frequency: 125ms (b) Frequency: 250ms (c) Frequency: 500ms (d) Frequency: 1000ms

Fig. 4 Total Packet Transmissions of Dumbbell Topology.

Table 4 RTO values of Grid Topology.

Packet frequency	PER= 0%			PER= 5%			PER= 10%		
	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA
125 ms	0.69	0.57	0.45	0.78	0.58	0.50	2.64	1.99	0.83
250 ms	1.79	0.98	0.62	2.09	1.69	1.47	5.86	4.11	1.97
500 ms	2.39	1.48	1.02	2.93	2.84	2.60	5.02	4.03	2.65
1000 ms	2.44	2.14	1.56	4.01	3.96	3.50	5.58	5.49	3.91

Table 5 RTO values of Dumbbell Topology.

Packet frequency	PER= 0%			PER= 5%			PER= 10%		
	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA
125 ms	1.60	1.26	0.80	3.13	1.96	1.26	3.88	1.93	1.57
250 ms	3.18	1.31	1.05	3.75	2.99	2.37	4.03	3.78	2.29
500 ms	3.24	1.79	1.40	4.57	4.34	3.42	6.90	6.21	4.32
1000 ms	3.53	2.72	2.16	5.58	5.24	4.29	9.29	7.87	7.04

and Flexi COCOA in Grid topology and Dumbbell topology are reported in Tables 6 and 7. Referring the Tables 6 and 7, we can realize that in general GSRTC uses larger weight values than Flexi COCOA. Larger weights of GSRTC make aggressive reduction of RTO compared to Flexi COCOA and results in spurious transmissions. We report the number of strong RTT samples obtained by GSRTC and Flexi COCOA in Tables 8 and 9. GSRTC achieves a smaller number of strong RTT samples than Flexi COCOA and thus we confirm spurious transmissions of GSRTC. Correspondingly, RTO

Table 6 Weight values in Grid Topology.

Packet frequency	PER=0%		PER=5%		PER=10%	
	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA
125 ms	0.95	0.93	0.89	0.86	0.83	0.77
250 ms	0.92	0.92	0.88	0.84	0.81	0.76
500 ms	0.91	0.88	0.85	0.81	0.80	0.76
1000 ms	0.87	0.83	0.83	0.79	0.78	0.75

Table 7 Weight values in Dumbbell Topology.

Packet frequency	PER=0%		PER=5%		PER=10%	
	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA
125 ms	0.92	0.92	0.87	0.84	0.82	0.77
250 ms	0.90	0.91	0.85	0.80	0.80	0.76
500 ms	0.88	0.88	0.83	0.76	0.79	0.73
1000 ms	0.83	0.82	0.82	0.71	0.77	0.68

Table 8 Strong RTT samples in Grid Topology.

Packet frequency	PER=0%		PER=5%		PER=10%	
	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA
125 ms	3798	4323	2740	2935	1520	2000
250 ms	2931	3516	1502	1654	469	920
500 ms	2387	2803	851	916	419	526
1000 ms	1572	2224	447	640	285	311

Table 9 Strong RTT samples in Dumbbell Topology.

Packet frequency	PER=0%		PER=5%		PER=10%	
	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA	GSRTC	Flexi COCOA
125 ms	3923	5569	2487	3240	1877	2519
250 ms	3562	4908	1660	2053	1015	1348
500 ms	3180	3944	1176	1513	643	807
1000 ms	2035	2318	991	1145	414	487

values of GSRTC are greater than that of Flexi COCOA. Smaller RTO values can result in spurious transmissions if the RTO convergence approach is aggressive. Similarly, if the weight reduction is aggressive, smaller weights may reduce the speed of convergence, adversely affecting transmission delay. Therefore, any method which results in smaller weight or smaller RTO doesn't always imply an enhancement over COCOA. Furthermore, optimal values for weight and RTO may be unreasonable in this context, as packet loss is not always caused by congestion. The cautious adaptation of Flexi COCOA helps prevent aggressive changes in weight and RTO, potentially enhancing the performance of COCOA in comparison to GSRTC.

Both in Grid and in Dumbbell topologies, total packet transmissions tend to decrease with higher PER or lower packet transmission frequency. RTO of COCOA tends to increase when PER increases or packet transmission frequency decreases. For every packet transmission frequency, as the PER increases, the impact of weak RTT samples increases, increasing RTO. For every PER, when the network is lightly loaded, sequence of similar small strong RTT samples is accumulated. This causes the impact of RTVAR in Eq. (3) to vanish which consequently leads to the generation of small RTO and sporadic retransmissions, and such retransmissions increase RTO. The weight value in Flexi COCOA and GSRTC are resulted from strong RTT samples and therefore also gets influenced similarly.

The tendency of RTO values in Flexi COCOA and GSRTC is also the same as that of COCOA for the same reasons explained. However, the loss of packets is not uniformly distributed. GSRTC increases weight value faster whenever possible and becomes aggressive. As a result, higher weight value and higher RTO are observed in GSRTC than in Flexi COCOA. On the contrary, Flexi COCOA makes gradual changes to weight value and reduces RTO better than GSRTC. Consequently, total packet transmissions of Flexi COCOA are larger than that of GSRTC and COCOA in the context of higher PER or lower packet transmission

Table 10 Retransmission effort of Grid Topology.

Packet frequency	PER= 0%			PER= 5%			PER= 10%		
	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA
125 ms	0.10	0.10	0.08	0.18	0.16	0.16	0.35	0.27	0.27
250 ms	0.17	0.14	0.09	0.21	0.19	0.20	0.44	0.38	0.31
500 ms	0.23	0.15	0.14	0.28	0.26	0.26	0.42	0.36	0.34
1000 ms	0.25	0.23	0.22	0.32	0.30	0.29	0.44	0.42	0.38

Table 11 Retransmission effort of Dumbbell Topology.

Packet frequency	PER= 0%			PER= 5%			PER= 10%		
	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA	COCOA	GSRTC	Flexi COCOA
125 ms	0.19	0.15	0.10	0.28	0.21	0.20	0.43	0.32	0.29
250 ms	0.23	0.18	0.12	0.31	0.28	0.26	0.38	0.39	0.33
500 ms	0.24	0.19	0.15	0.34	0.33	0.31	0.46	0.44	0.38
1000 ms	0.27	0.26	0.24	0.37	0.35	0.35	0.50	0.50	0.46

frequency.

From Figs. 3 and 4, we can observe that Dumbbell topology has a greater number of packet transmissions compared to Grid topology. The positioning of nodes varies between the Grid topology with 9 nodes and the dumbbell topology with 21 nodes as shown in Fig. 2. In both topologies, each node has direct neighbors 10 meters away. The number of direct neighbors and the number of nodes simultaneously competing for the radio channel are more in Dumbbell topology than in Grid topology. The radio link connecting two halves of the Dumbbell is a bottleneck, causing an increase in retransmissions if not shared fairly. Consequently, Dumbbell topology results in more weak RTT samples than Grid topology. This is evident from Tables 6 and 7 that the weight values of Dumbbell are generally less than that of Grid. As a result, the RTO of Dumbbell topology is greater than that of Grid topology as in Tables 4 and 5. However, packet loss varies among nodes of Dumbbell topology due to the randomization of algorithms implemented across different layers of the protocol stack.

The Dumbbell topology provides a benefit over the Grid topology regarding the length of RTT. The length of route between CoAP client and CoAP server can lead to the length of RTT. In Grid topology, the number of hops between them varies, since many combinations of links for the connection of client and server nodes are possible. Hence, RTT varies accordingly. In Dumbbell topology, clients on half circles are two hops away from the CoAP server. Therefore, the length of strong RTT sample is lesser in Dumbbell topology than in Grid topology. Nodes experiencing higher losses back off from transmission for longer durations. Nodes experiencing lesser losses exploit the back off duration of nodes with higher losses to successfully deliver packets to the server, taking advantage of the length of RTT in the Dumbbell topology. Consequently, the total packet transmissions in the Dumbbell topology are higher than Grid topology.

GSRTC and Flexi COCOA further improves the delay in transmission with the help of adaptive weight values by providing more weightage to current strong RTT. However, Flexi COCOA demonstrates a competent refining of weight value than GSRTC. Performance of GSRTC is affected by the impulsive reset of weight to 0.5 after a weak RTT is computed. Flexi COCOA manages RTO computation better than COCOA and GSRTC even with infrequent weak RTT samples. Therefore, Flexi COCOA exhibits better performance than COCOA and GSRTC in total packet transmissions. Upgrading the bandwidth of the bottleneck link may help alleviate congestion in a Dumbbell topology. However, the limited memory and processing speed of constrained nodes make discussions on higher bandwidth utilization irrelevant.

5.2.2 Retransmission Effort

Retransmission in constrained network is costlier. Therefore, retransmission is a crucial evaluation criterion. We analyze the context of retransmission using retransmission effort metric. The retransmission effort of the client is defined as the average number of retransmissions attempted to deliver each data packet to the server. The ratio between total retransmissions and total data packets is computed as retransmission effort. Tables 10 and 11 show the results obtained. Flexi COCOA achieves lower retransmission effort than COCOA and GSRTC in all the experiments we conducted. The tendency of retransmission effort values is like that of RTO values. The reasons are same as already explained for RTO. For the sake of brevity, we do not repeat the explanation.

We dissect the RTO_{init} and RTO_{new} to understand the result further. The number of message transmissions that start with large RTO_{new} values in COCOA is higher than that of Flexi COCOA. However, when strong RTT is obtained in COCOA, RTO_{new} coverages to RTT uniformly based on the constant weight value. On the other hand, Flexi COCOA breaks the uniformity in converging RTO_{new} to RTT. The weight value of Flexi COCOA is never a pre-fixed value and does not exhibit monotonic increase or decrease. Therefore,

RTO_{new} in Flexi COCOA, responds quickly to network dynamics than COCOA and derives sensible RTO_{init} values. As a result, retransmission effort in Flexi COCOA is less than COCOA. When weak RTT is computed, GSRTC inherits *Fullbackoff2* [17] and controls retransmissions. Using *Fullbackoff2*, a new transmission after a weak RTT sample makes uses maximum of back off RTO or RTO_{new} as RTO_{new}. Hence, GSRTC achieves smaller retransmission effort than COCOA by reusing larger RTO value from previous transmission. However, the rapid increase of weight value using strong RTT samples makes the retransmission effort of GSRTC larger than that of Flexi COCOA.

Tables 10 and 11 show that retransmission effort of Dumbbell topology is greater than that of Grid topology. The number of nodes simultaneously competing for the radio channel is more in Dumbbell topology than in Grid topology. Therefore, retransmission due to packet collision increases. The radio link between two halves of the Dumbbell is also a critical point. When the bandwidth offered by this link does not suffice, retransmission increases. Such characteristics of Dumbbell topology give rise to more retransmission effort when compared to Grid.

Using smaller RTO values, the number of transmissions is expected to increase. However, if the reduction in RTO is inefficient, retransmissions increase due to packet loss. An efficient RTO needs to be always as much associated with network dynamics as possible. Flexi COCOA is always adaptive to changes in the network signaled by strong RTT samples. Therefore, RTT samples contribute sensitively in Flexi COCOA compared to COCOA and GSRTC. Retransmission effort values of Flexi COCOA prove the reduction of RTO is efficient.

5.3 Discussion

Flexi COCOA outperforms COCOA and GSRTC in all performance metrics of our experiments. We analyze the result in three aspects: packet transmission frequency, PER and topology. The general trend we observed in the performance evaluation is that Flexi COCOA exhibits the best performance for high frequency transmissions in an error free environment.

In an error free network, the probability of obtaining strong RTT is larger. Therefore, the weight value becomes proportionally higher, and the reduction of RTO is faster. Flexi COCOA reduces transmission delay using a gradually changing weight value against the rapid change of GSRTC and gives the best result in our experiments. When the packet transmission frequency reduces, a flexible weight value still reduces the delay in transmission. However, nodes also must wait until the frequency timer expires. Therefore, high frequency transmission shows more change than low frequency transmissions.

CoAP expects that the constrained nodes may work in the lossy network. Therefore, we introduced PER in the network to understand the performance of Flexi COCOA in lossy network. In lossy network, we attempt to utilize all chances of obtaining strong RTT and improve the weight parameter whenever it is possible. Even with PER, GSRTC is found to be more aggressive than Flexi COCOA. From the results, we prove that Flexi COCOA is more reliable than COCOA and GSRTC, for applications even to work in the presence of errors. Flexi COCOA solely depends on strong RTT samples to improve the efficiency of RTO computation using better weight value than COCOA and GSRTC. When PER increases, strong RTT decreases and accordingly declines the weight value. This poses a challenge for Flexi COCOA to deliver better RTO than COCOA. Reduction of weight value as the PER increases is visible in Tables 6 and 7. When the number of weak RTT increases more than that of strong RTT, the network is extremely lossy. A conservative mechanism is beneficial over aggressive mechanism here. Hence Flexi COCOA contributes only the least.

We evaluated Grid topology and Dumbbell topology to understand how positioning of nodes affect the performance of Flexi COCOA. When there are more direct neighbors, the probability of simultaneous access of the same link increases and which in turn results in weak RTT samples. Therefore, Dumbbell topology shows larger RTO and more retransmission effort compared to Grid. When a greater number of nodes are closer to the server, the length of strong RTT becomes lesser and the number of transmissions can be improved. As a result, Dumbbell topology has more packet transmissions compared to Grid topology. Thus, the positioning of nodes can impact the margin of performance differences that Flexi COCOA can offer against COCOA and GSRTC.

6. Conclusion and Future Work

In this paper, we proposed Flexi COCOA to enhance the network sensitivity of COCOA using flexible weight value. We successfully addressed the challenge of generating network sensitive RTO with the help of flexible weight value of strong RTO computation. Flexi COCOA exhibits desired performance with change in frequency or PER. Results indicate that the proposed method can improve the QoS of IoT monitoring applications. Flexi COCOA can be easily integrated into COCOA. The proposed method utilizes monitoring of already available measures in COCOA. Therefore, Flexi COCOA does not increase computation and communication overheads when compared to COCOA.

As future work, we will evaluate the perspective of traffic bursts, fairness, and energy consumption.

References

[1] K. Ashton et al., “That ‘internet of things’ thing,” RFID Journal, vol.22, no.7, pp.97–114, 2009.
[2] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” Computer Networks, vol.54, no.15, pp.2787–2805, 2010.
[3] Z. Shelby, K. Hartke, and C. Bormann, “The constrained application protocol (CoAP),” RFC 7252, 2014.
[4] T. Winter, P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J.P. Vasseur, and R. Alexander, “RPL: IPv6 routing

protocolfor low-power and lossy networks," Technical Report, RFC 6550, 2012.
[5] N. Kushalnagar, G. Montenegro, and C. Schumacher, "IPv6 overlow-power wireless personal area networks (6LoWPANs): Overview, assumptions, problem statement, and goals," RFC 4919, 2007.
[6] D. Pandey and V. Kushwaha, "An exploratory study of congestion control techniques in wireless sensor networks," Computer Communications, vol.157, pp.257–283, 2020.
[7] L. Eggert and G. Fairhurst, "Unicast UDP usage guidelines for application designers," RFC 5405, 2008.
[8] L. Eggert, "Congestion control for the constrained application protocol (CoAP)," draft-eggert-core-congestion-control-01 (work in progress), 2011.
[9] A. Betzler, C. Gomez, I. Demirkol, and J. Paradells, "CoAP congestion control for the internet of things," IEEE Commun. Mag., vol.54, no.7, pp.154–160, 2016.
[10] A. Betzler, C. Gomez, I. Demirkol, and J. Paradells, "CoCoA+: An advanced congestion control mechanism for CoAP," Ad Hoc Networks, vol.33, pp.126–139, 2015.
[11] P. Karn and C. Partridge, "Improving round-trip time estimates in reliable transport protocols," ACM SIGCOMM Computer Communication Review, vol.17, no.5, pp.2–7, 1987.
[12] V. Paxson, M. Allman, J. Chu, and M. Sargent, "Computing TCP's retransmission timer," RFC 6298, 2011.
[13] E. Ancillotti and R. Bruno, "Comparison of CoAP and CoCoA+ congestion control mechanisms for different IoT application scenarios," IEEE Symposium on Computers and Communications (ISCC), pp.1186–1192, 2017.
[14] R. Bhalerao, S.S. Subramanian, and J. Pasquale, "An analysis and improvement of congestion control in the CoAP internet-of-things protocol," IEEE Annual Consumer Communications & Networking Conference (CCNC), pp.889–894, 2016.
[15] H. Meng, H. HongBing, and L. WeiPing, "An adaptive congestion control algorithm with CoAP for the Internet of Thing," Int. J. Comput. Tech., vol.4, pp.46–51, 2017.
[16] I. Jarvinen, I. Raitahila, Z. Cao, and M. Kojo, "Fasor retransmission timeout and congestion control mechanism for CoAP," IEEE Global Communications Conference (GLOBECOM), pp.1–7, 2018.
[17] I. Järvinen, I. Raitahila, Z. Cao, and M. Kojo, "Is CoAP congestion safe?," Proc. Applied Networking Research Workshop, pp.43–49, 2018.
[18] S. Bolettieri, G. Tanganelli, C. Vallati, and E. Mingozzi, "pCoCoA: A precise congestion control algorithm for CoAP," Ad Hoc Networks, vol.80, pp.116–129, 2018.
[19] V. Rathod, N. Jeppu, S. Sastry, S. Singala, and M.P. Tahiliani, "CoCoA++: Delay gradient based congestion control for Internet of Things," Future Generation Computer Systems, vol.100, pp.1053–1072, 2019.
[20] C. Suwannapong and C. Khunboa, "Congestion control in CoAP observe group communication," Sensors, vol.19, no.15, p.3433, 2019.
[21] S. Deshmukh and V.T. Raisinghani, "Adcocoa-adaptive congestion control algorithm for CoAP," IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp.1–7, 2020.
[22] G.A. Akpakwu, G.P. Hancke, and A.M. Abu-Mahfouz, "CACC: Context-aware congestion control approach for lightweight CoAP/UDP-based Internet of Things traffic," Trans. Emerging Telecommunications Technologies, vol.31, no.2, p.e3822, 2020.
[23] R.K. Yadav, N. Singh, and P. Piyush, "Genetic CoCoA++: Genetic algorithm based congestion control in CoAP," IEEE International Conference on Intelligent Computing and Control Systems (ICICCS), pp.808–813, 2020.
[24] A.K. Demir and F. Abut, "mlCoCoA: A machine learning-based congestion control for CoAP," Turkish Journal of Electrical Engineering and Computer Sciences, vol.28, no.5, pp.2863–2882, 2020.
[25] D.H. Hoang and T.T.D. Le, "RCOAP: A rate control scheme for reliable bursty data transfer in IoT networks,"IEEE Access, vol.9, pp.169281–169298, 2021.
[26] V. Rathod and M.P. Tahiliani, "Geometric series based effective RTO estimation technique for CoCoA," Ad Hoc Networks, vol.130, p.102801, 2022
[27] G.A. Akpakwu, G.P. Hancke, and A.M. Abu-Mahfouz, "An optimization-based congestion control for constrained application protocol," International Journal of Network Management, vol.32, no.1, p.e2178, 2022.
[28] F. Righetti, C. Vallati, D. Rasla, and G. Anastasi, "Investigating the CoAP congestion control strategies for 6tisch-based IoT networks," IEEE Access, vol.11, pp.11054–11065, 2023.
[29] P.K. Donta, S.N. Srirama, T. Amgoth, and C.S.R. Annavarapu, "iCoCoA: Intelligent congestion control algorithm for CoAP using deep reinforcement learning," J. Ambient Intell. Human. Comput., vol.14, no.3, pp.2951–2966, 2023.
[30] "Contiki," https://github.com/contiki-os/contiki
[31] "Contiki tutorial," https://anrg.usc.edu/contiki/index.php/Contiki_tutorials
[32] "Tmote sky," https://telosbsensors.wordpress.com/
[33] "Zolertia Z1 mote," https://github.com/Zolertia/Resources/wiki/The-Z1-mote

Archana K. Rajan received B.Tech degree in Information Technology from Kerala University, India in 2006. She received M.Tech degree in Computer Science and Engineering from Amrita Vishwa Vidyapeetham, India in 2013. She is currently a Ph.D. student in the Department of Information and Communication Sciences at Sophia University, Japan. Her current research interests include IoT and congestion control.

Masaki Bandai received the B.E. and M.E. degrees in Electrical Engineering and the Ph.D. degree in Information and Computer Science from Keio University, Japan in 1996, 1998, and 2004, respectively. From 2004 to 2010, he was an assistant professor at Shizuoka University, Japan. He is currently a professor in the Department of Information and Communication Sciences at Sophia University, Japan. His current research interests include computer networks, network computing, and applications.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

Model for Controller Assignment and Placement to Minimize Migration Blackout Time with Load-Balancing Platform in Software-Defined Network

Shinji NODA[†], ***Student Member***, **Takehiro SATO**[†a)], ***Member***, ***and*** **Eiji OKI**[†], ***Fellow***

SUMMARY A software-defined network (SDN) is a network that the centralized SDN controller controls multiple SDN switches. Load-balancing platforms can realize the distribution of the load of the switches between multiple controllers. The platforms allow controller processing capacity to be used efficiently. When the assignment between switches and controllers and the controller placement are changed, migration blackout time that the controller temporarily stops processing messages can occur. The migration blackout time can result in failure to meet delay requirements between switches and controllers. This paper proposes a model that determines the controller assignment and placement while minimizing the migration blackout time with the load-balancing platform. The proposed model can be used when the controllers in the network are overloaded and the controller assignment and placement need to be changed. We formulate the proposed model as a mixed-integer second-order cone programming problem. We develop a migration procedure used in the proposed model. In the procedure, each switch can be controlled by multiple controllers with a load-balancing platform. The load-balancing platform allows status messages sent from a switch to be sent to multiple controllers. This allows status messages sent from the switches to be processed in order and the migration blackout can be avoided. The proposed model is compared with a baseline model based on the previous works. In the baseline model, the migration blackout time always occurs when the controller assignment and placement are changed. Numerical results show that the migration blackout time in the proposed model becomes smaller than that in the baseline model. The results also show that the number of controllers placed in the proposed model is smaller than that in the baseline model.

key words: *software-defined network, controller placement, load balancing, switch migration, migration blackout*

1. Introduction

In software-defined networks (SDNs), control devices are separated from forwarding devices and a logically-centralized controller can control multiple forwarding devices. This enables flexible and centralized network management [1], [2].

Deploying multiple controllers alleviates problems in the SDN. For example, if the amount of load assigned to a controller is likely to exceed its processing capacity, the controller load can be distributed to another controller by reassigning switches. This realizes load-balancing between controllers. In addition, if a controller fails, switches connected to the controller can be assigned to another controller. This realizes network management with fault tolerance.

One of the flow setup methods, which is adopted to realize SDN, is OpenFlow [3]. Packets arriving at a switch are processed based on the procedures defined by OpenFlow. When a packet arrives, the switch sends a packet-in message to a controller. The controller decides what action the packet should take and sends the result to the switch as a packet-out message. This procedure incurs propagation delay, which is the time it takes for a message to be propagated between the switch and the controller, and the sojourn time, which is the time it takes for the controller to decide the action that the packet takes. The sojourn time is expressed as the sum of two types of time. One is the time spent waiting to be processed at the controller. The other is the time it takes to be processed at the controller. The sojourn time depends on the amount of messages arriving at the controller. As the controller load becomes larger, the sojourn time on the controller increases. The sum of the propagation delay and the sojourn time must not exceed the delay requirement from the network management perspective.

There are several platforms for distributing the load of a switch to multiple controllers. FlowVisor [4] is a platform that creates a network slice for each controller in the SDN. By using these slices, multiple controllers can control a single switch. FlowVisor acts as a proxy between a switch and controllers, and OpenFlow messages that are exchanged between a switch and controllers are passed through FlowVisor. OpenVirteX [5] is a network virtualization platform that spawns virtual networks that have arbitrary topology and addressing schemes. This makes it possible for each switch to be controlled by multiple controllers. HyperFlex [6] is a hypervisor architecture that virtualizes the control plane of SDNs, and makes it possible for multiple controllers to control a switch by isolating network resources. By using such platforms, load-balancing in the SDNs can be achieved. This allows the processing capacity of each controller to be used efficiently.

In a real network environment, the number of packets arriving at a switch varies. Therefore, the efficiency of the assignment between switches and controllers and the controller placement that are once determined will be lost, which causes violation of the delay requirement between switches and controllers. In this case, it is needed to change the controller assignment and placement. In order to guarantee the order in which messages are processed during the reassign-

Manuscript received March 5, 2024.
Manuscript revised May 1, 2024.
Manuscript publicized June 28, 2024.

[†]Graduate School of Informatics, Kyoto University, Kyoto-shi, 606-8501 Japan.
a) E-mail: takehiro.sato@i.kyoto-u.ac.jp

DOI: 10.23919/transcom.2024EBP3044

ment operation, controllers need to temporarily stop processing messages during the reassignment operation. This time is called the migration blackout [7]. During this blackout period, the controller cannot process any messages arriving from the switch; it is desirable that the migration blackout should not occur.

There is a study that presented a migration procedure in the environment where the amount of packets on the network varies. Dixit et al. [7] developed an architecture to control the controller load by changing the assignment between switches and controllers. In this work [7], a protocol for migrating switch information based on the OpenFlow standard was developed. The protocol changes the assignment to realize controller load balancing. The work [7] noted that, in the protocol, there is a time that the controller needs to stop processing. This is the migration blackout time.

There are several studies that determine controller assignment while considering the migration blackout time. Xu et al. [8] presented an algorithm to achieve load balancing among controllers. In the algorithm presented in [8], switches to be migrated are selected so that the network is not disjointed due to the migration blackout. Yue et al. [9] presented a scheme that minimizes the end-to-end delay between switches and controllers while considering the migration blackout time and load-balancing. In the scheme, the migration blackout time inevitably occurs when the switch is reassigned to another controller. Min et al. [10] presented heuristic solutions that aim to minimize the queuing and processing times of the requests from switches at the controllers and balance the controller loads while incorporating the switch migration cost. The solutions consider the number of packets affected by the switch migration procedure as the switch migration cost when the controller assignment and placement are determined.

While previous studies [8]–[10] consider the inability to process packets arriving at the switch during the migration blackout, they do not address preventing the migration blackout. When the migration blackout occurs, the response to packet-in messages generated by a switch is delayed, which can result in failure to meet the delay requirements between switches and controllers. In the previous study [8], the period of the migration blackout was reported to increase up to 370 ms. On the other hand, the time required for flow setup between switches and controllers is set to 200 ms in [11]; the impact of the migration blackout on the delay requirement cannot be ignored. It is required to balance the controller load while suppressing the migration blackout.

Migration blackout can be prevented by using platforms for distributing the load of a switch such as FlowVisor. In the previous studies [8]–[10], the migration blackout occurs in order to maintain the processing order of messages arriving at switches when the assignment between switches and controllers is changed. A typical example of the messages is the status message from switches, which is sent to controllers to inform the state of the switches. If we apply the load balancing platform, a single switch can be connected to multiple controllers simultaneously, and the status messages from the switch are sent to the multiple controllers that are connected to the switch. This can allow messages to be processed in order without causing the migration blackout when changing the controller assignment.

Note that the migration blackout can be completely avoided if controllers are placed in all nodes in the network and each switch is connected to all of the controllers. However, in the actual network management, the number of switches that can be controlled by one controller is limited since the controller can only process a limited amount of messages sent from switches. In addition, as the number of placed controllers increases, the communication overhead between controllers increases, which wastes network capacity. Therefore, the network management that minimizes the migration blackout time with a limited number of controllers is needed.

Our previous works in [12], [13] studied the controller assignment and placement problem in an environment with a load-balancing platform. The work in [12] developed a controller placement model that deals with controller failures. The model determines the controller assignment and placement so that the switch load can be distributed to active controllers in the event of controller failures. The work in [13] extended the model in [12] by taking into account the sojourn time at each controller. The model presented in [13] expresses the sojourn time based on the queuing theory; the sojourn time changes depending on the amount of load arriving at each controller. These works focused on scenarios of controller failures and did not consider the migration blackout time.

This paper proposes a model that determines the controller assignment and placement to minimize the migration blackout time with a load-balancing platform. Different from our previous studies [12], [13], the proposed model considers the migration blackout time. The proposed model is used when the controller assignment and placement need to be changed due to the overload on the controller. Along with the proposed model, the contributions of this paper include the following:

- We develop a migration procedure used in the proposed model. By adopting a load-balancing platform, each switch can be controlled by multiple controllers. The load-balancing platform can determine the ratio of the amount of load assigned from a switch to controllers. Each controller does not need to stop processing messages when the connections between the switch and the controllers do not change and the ratio of the amount of load assigned from the switch to the controllers is changed. This is because the load-balancing platform allows messages sent from a switch to be sent to multiple controllers and the controllers process messages in proper order; the processing order is preserved without setting the migration blackout time. Our migration procedure can deal with this case. Different from the previous works that dealt with the switch migration [8]–[10], this procedure migrates switches with

the load-balancing platform.
- We formulate the proposed model as an optimization problem. In the optimization problem, a certain constraint is nonlinear to handle the time to process packets queued at a controller. We transform the optimization problem and express it as a mixed-integer second-order cone programming problem.
- We evaluate the performance of the proposed model in both static and dynamic scenarios. In the static scenario, where the initial controller placement and the assignment between switches and controllers are given, the migration blackout time that occurs when the controller assignment and placement are changed is evaluated. In the dynamic scenario, the migration blackout time is evaluated under the condition that the amount of load generated by the switch varies.

The rest of this paper is organized as follows. Section 2 describes the related works. Section 3 describes the migration procedure assumed in the proposed model. Section 4 presents the proposed model. Section 5 presents the baseline model. Section 6 shows numerical results. Finally, we conclude this paper in Sect. 7.

2. Related Works

There have been several works that deal with the controller placement in the environment where the load generated on the switches varies. Hegde et al. [14] presented controller placement algorithms for an SDN, where the edge and core of the network are separated. This work [14] also discussed the switch migration procedure for the transfer of a switch from one controller to another in order not to disrupt ongoing connections. Mouawad et al. [15] addressed the controller placement problem and dealt with network load variation using a dynamic switch migration algorithm. The work aims at finding an optimal dynamic SDN controller placement that reduces controller-switch latency, inter-controller latency, and controller load. Liu et al. [16] presented a dynamic matching algorithm that achieves controller load balance. The algorithm is implemented for maximizing the control resource utilization, with the condition that the controller is not overloaded. The previous works [14]–[16] do not consider the migration blackout time in their presented algorithms. Different from [14]–[16], our work considers the migration blackout time when the controller assignment and placement are changed.

Some works have dealt with the switch migration between controllers in the environment where the amount of packets in the network varies. Zadeh et al. [17] presented a general framework for converting an existing switch migration protocol to one in which multiple controllers are simultaneously involved. This framework suppresses the worst-case migration latency. Yue et al. [9] presented a scheme that minimizes end-to-end delay between switches and controllers while considering the migration blackout time and load balancing. Min et al. [10] presented heuristic based solutions to minimize the latency, balances the controller loads, and reduces the switch migration cost all at once. The solution is designed in order to solve the switch migration problem comprehensively and effectively in dynamic Internet of Things networks. Different from [9], [10], [17], our work minimizes the migration blackout time with the load-balancing platform.

Dixit et al. [7] presented a migration protocol based on OpenFlow to migrate the load on a controller to another controller. They assumed the situation where the controller load fluctuates dynamically and the load on one controller increases. The protocol is designed to satisfy *Serialization*, *Safety*, and *Liveness*. *Serialization* means that messages are processed in the order in which they are sent. If messages sent from a switch are not processed in the order, the controller's perception of the switch can be different from that actual state. For example, if a port on a switch goes down and then comes back up, the switch sends a port status down message followed by a port status up message. If the order of processing port status messages is reversed, the controller recognizes a port that is actually available as unavailable. *Safety* means preventing duplicate processing of messages; duplicate processing of packet-in messages may result in duplicate flow entries and cause failures in the database. *Liveness* means that packets arriving at switches are always processed by at least one controller to which they are connected. If no controllers are connected to the switch, packets arriving at the switch are not routed properly. In the migration protocol presented in the previous study [7], there is a time period when packets arriving at the switch cannot be processed, which is called the migration blackout time. The previous study [7] does not address the suppression of this period. Different from the work, our proposed model assumes the load-balancing platform in the migration procedure.

Xu et al. [8] presented an algorithm to solve the problem of updating allocations between switches and controllers in a dynamically changing network environment. The objective function of the problem is to minimize the processing load on the controller when directing packets generated at the switches to their destinations. If the load on a controller exceeds a threshold, the algorithm changes the controller assignment for the load balancing. The previous work [8] considers the migration blackout that occurs when the assignment between switches and controllers is changed. When the migration blackout occurs, messages arriving from the switch cannot be processed, resulting in network fragmentation. The algorithm selects the switch to be migrated in order to prevent network fragmentation due to migration. Different from the work [8], our model directly suppresses the migration blackout itself.

3. Migration Procedure

3.1 4-Phase Migration Protocol

Dixit et al. [7] presented a procedure to reassign a switch

from one controller to another controller while satisfying the three properties: *Serialization*, *Safety*, and *Liveness*. In the procedure, the assignment of the switch is changed from controller A to controller B in four phases. The initial role of controller A for the switch is *master* and the initial role of controller B is *slave*. In the first phase, the role of controller B for the switch is changed to *equal* to ensure *Liveness*. In the second phase, the controller that is primarily responsible for controlling the switch is changed from A to B to ensure *Safety*. In the third phase, messages accumulated at controller A are processed before controller B begins to control the switch to ensure *Serialization*. In the fourth phase, the role of controller A to the switch is changed to *slave* and then the role of controller B is changed to *master*.

3.2 Processing Messages with Load-Balancing Platform

When a switch is controlled by several controllers with the load-balancing platform, the platform allows multiple controllers to process messages sent from the switch. Messages that must be processed in the order in which they are sent from the switch to the controller are called dependent messages. Dependent messages are replicated at the load-balancing platform and processed by several controllers. An example of the dependent message is the port status message. This message informs the status of the ports on the switch to the controllers. Consider the case where a switch is connected to two controllers A and B using a load balancing platform. The load balancing platform acts as a proxy between the switch and the controllers. When a port status message is sent from the switch, the message is sent to both controllers A and B by replicating the message and changing the destination of the message. This procedure allows several controllers to have the same perception of the switch, and multiple controllers can process messages sent from the switch in a proper order.

3.3 Migration Procedure Used in Proposed Model

This section describes the message sequences of the migration procedure used in the proposed model that is shown in Sect. 4. We develop the message sequences based on the one presented in [7]. In the procedure, a load-balancing platform is applied and each switch can be controlled by multiple controllers.

There are four cases of reallocation between switches and controllers. We describe the exchange of messages between switches and controllers for each case below. We assume the case where a switch and controllers A and B exist, and the connection of the switch to the controllers is changed. The switches to be migrated are determined by monitoring the network based on the increase or decrease in the amount of packets generated and the amount of load that is assigned to the controllers.

3.3.1 Case where Connection between Switch and Controllers is Changed

In this case, the connection between the switch and controller A is deleted and a new connection between the switch and controller B is established. In this migration procedure, the role of controller A for the switch is changed from *master* to *slave*, and the role of controller B is changed from *slave* to *master*. Since controller B needs to receive information about the switch from controller A, a migration blackout occurs. The sequence diagram for the procedure with a load-balancing platform is shown in Fig. 1.

In the first phase, controller A sends a *Start migration* message to controller B. When controller B receives the message, it changes its role to *equal* to the switch. Controller B then sends a *Ready for migration* message to controller A. Messages between controllers A and B and the switch are adequately exchanged by the load-balancing platform.

In the second phase, controller A sends a *Request to update action list* message to the load balancing platform. The list of actions in the load balancing platform is updated and the destination of messages sent from the switch is changed. This action changes the controller that is responsible for control of the switch from A to B.

In the third phase, messages accumulated at controller A are processed before controller B begins to control the

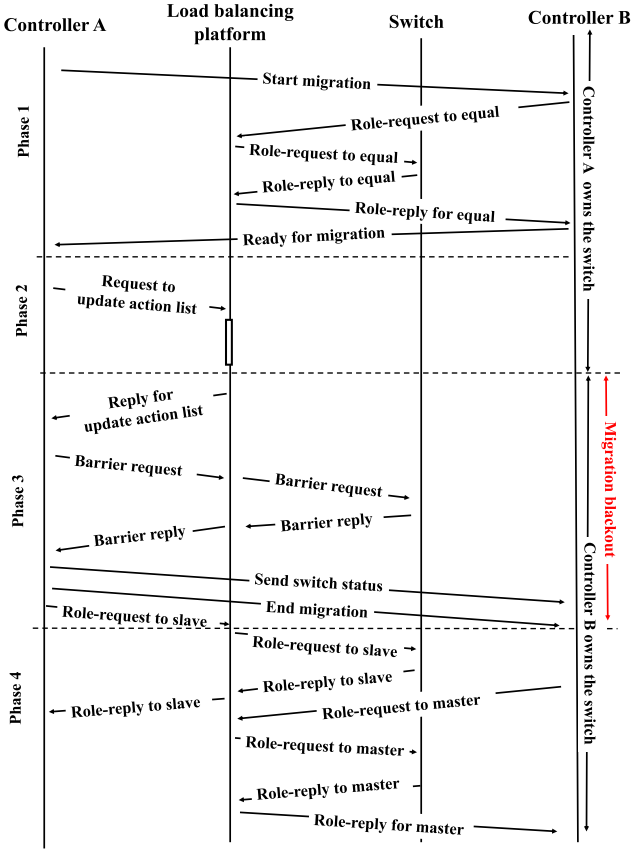


Fig. 1 Case where connection between switch and controllers is changed.

switch. Although controller B is now in control of the switch, it is not possible to install a flow entry on the switch. This is because there are still unprocessed messages on controller A. Controller B needs to wait to process the messages that have accumulated on controller A. Messages arriving at controller B from the switch are buffered. When the processing of messages accumulated in controller A is completed, controller A sends a *Barrier request* message to the switch, confirms that there are no unprocessed messages, and sends information on controller A to controller B. After that, controller A then sends an *End migration* message to controller B.

In the fourth phase, the role of controller A with respect to the switch is changed to *slave* and then the role of controller B is changed to *master*.

3.3.2 Case where Number of Connections between Switch and Controllers Increases

In this case, a new connection between the switch and controller B is established in a situation where the switch is connected to controller A. In this migration procedure, the role of controller A for the switch is not changed, and the role of controller B is changed from *slave* to *master*. Since controller B needs to receive information about the switch from controller A, the migration blackout occurs. The sequence diagram for the procedure with a load balancing platform is shown in Fig. 2.

The procedure from the first to third phases is the same as that in Fig. 1.

In the fourth phase, the role of controller A is not changed and the role of controller B is changed to *master*.

3.3.3 Case where Number of Connections between Switch and Controllers Decreases

In this case, the connection between the switch and controller B is deleted in a situation where the switch is connected to controllers A and B. Controllers A and B already have information about the switch, so there is no need to transfer information between the controllers; the migration blackout does not occur. In this procedure, the role of controller A with respect to the switch is not changed, and the role of controller B is changed from *master* to *slave*. The sequence diagram for the procedure with a load balancing platform is shown in Fig. 3.

In the first phase, controller A sends a *Start migration* message to controller B. Controller B then sends a *Ready for migration* message to controller A.

The procedure in the second phase is the same as those in Figs. 1 and 2.

In the third phase, the messages that have accumulated in controller B before the connection status of the switch is changed are processed. After the processing is completed, a *Barrier request* message is sent to the switch, and with its response, it is confirmed that there are no unprocessed messages. After that, the *End migration* message is sent to

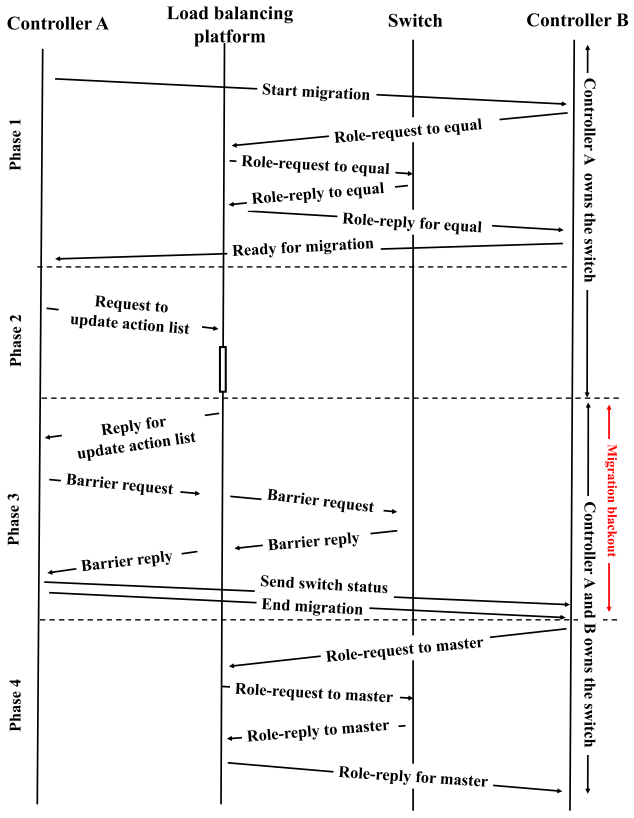


Fig. 2 Case where number of connections between switch and controllers increases.

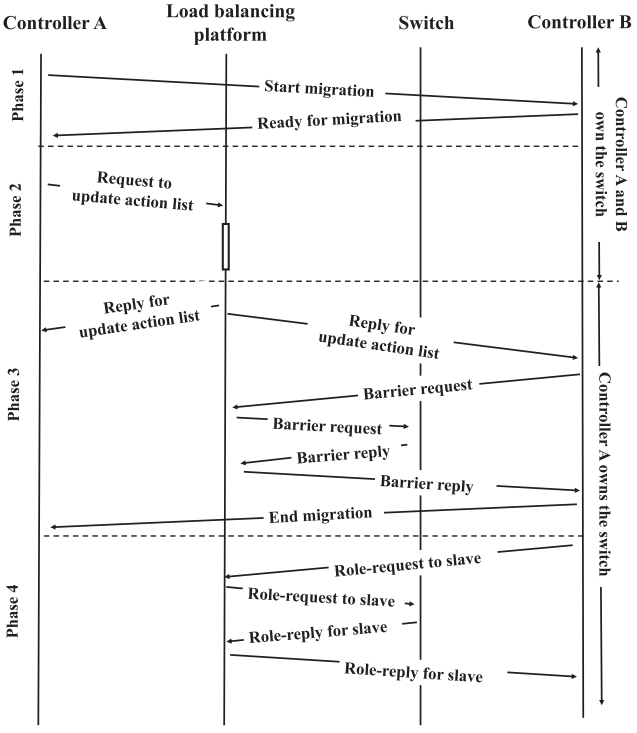


Fig. 3 Case where number of connections between switch and controllers decreases.

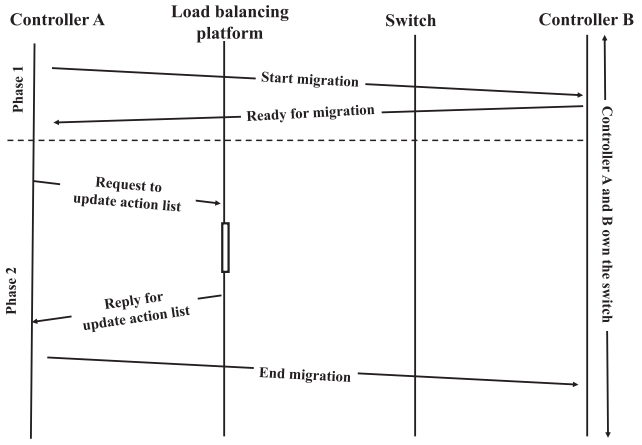


Fig. 4 Case where connections between switch and controllers do not change, and ratio of amount of load assigned from switch to controllers is changed.

controller B to inform that the migration procedure has been completed.

In the fourth phase, the role of controller B is changed to *slave*.

3.3.4 Case where Ratio of Amount of Load Assigned from Switch to Controllers is Changed

In this case, the switch is connected to controllers A and B, and the amount of load allocated from the switch to the controllers is changed. The controllers already have information about the switch, so there is no need to transfer information between controllers; the migration blackout does not occur. The roles of controllers A and B with respect to the switch are not changed in this procedure. The sequence diagram for the above procedure with a load balancing platform is shown in Fig. 4.

The first phase is the same as that in Fig. 3.

In the second phase, in the same way as those in Figs. 1, 2, and 3, the list of actions in the load balancing platform is updated and the destination of the messages sent from the switch is changed. The procedure is then completed by sending an *End migration* message to controller B.

4. Proposed Model

4.1 Overview

The proposed model determines the assignment between switches and controllers or the deployment of new controllers. The proposed model is based on the migration procedure that is described in Sect. 3.3. We assume that the proposed model is used when a certain percentage of the controllers' maximum processing capacity is used. The current assignment between switches and controllers, the current placement of controllers, and the amount of load generated on the switches are given. The proposed model determines the new assignment and the controller placement in order to minimize the migration blackout time. In addition, when deploying a new controller, the proposed model determines which switch is connected to the controller based on the migration blackout time. The proposed model considers the delay requirement between switches and controllers and the capacity constraint of each controller.

Table 1 Notations of sets and parameters used in this paper.

Given set and parameter	Definition
S	Set of switches.
C	Set of nodes where controllers can be deployed.
$\tilde{x}_{i,j}$	Ratio of the amount of load assigned from switch $i \in S$ to controller placed in node $j \in C$ to the amount of load generated on switch $i \in S$ before updating controller assignment.
$\tilde{n}_{i,j}$	Constants indicating whether switch $i \in S$ and controller placed in node $j \in C$ are connected before updating controller assignment.
$\tilde{c}_j$	Constants indicating whether controller is located at node $j \in C$ before updating controller assignment.
t_k^{mig}	Migration blackout time occurring at controller placed in node $j \in C$.
α_j	Packet arrival rate at controller placed in node $j \in C$.
ζ_j	Packet processing rate at controller placed in node $j \in C$.
d_i	Amount of packets arriving at switch $i \in S$.
$t_{i,j}^{\mathrm{short}}$	Minimum propagation delay between switch $i \in S$ and controller placed in node $j \in C$.
t^{lim}	Upper bound of the delay between switches and controllers.
$t_{j,k}^{\mathrm{con}}$	Delay in migration between controllers placed in nodes $j, k \in C$.
t_j^{proc}	Time required to process packets accumulated on controller placed in node $j \in C$ when the information of switches is migrated at the controller.
ρ^{delay}	Threshold for delay limit.
ρ^{capa}	Threshold for capacity limit of controllers.
ρ^{limit}	Maximum usable processing capacity of each controller.
c^{max}	Maximum number of placed controllers.

4.2 Notations

The notations of sets and parameters used in this paper are summarized in Table 1. The main sets and parameters are as follows. Let S denote a set of switches, and C denote a set of nodes that controllers can be deployed. Let $\tilde{x}_{i,j}$ denote the ratio of the amount of load assigned from switch $i \in S$ to a controller placed in node $j \in C$ to the amount of load generated on switch $i \in S$ before updating the controller assignment. Let $\tilde{n}_{i,j}$ denote an initial assignment from switch $i \in S$ to a controller placed in node $j \in C$. Let $\tilde{c}_j$ denote an

initial controller assignment placed in node $j \in C$. Let α_j and ζ_j denote the packet arrival rate and the packet processing rate of a controller placed in node $j \in C$, respectively. Let $t^{\mathrm{con}}_{j,k}$ denote the delay in migration between controllers placed in nodes $j, k \in C$. Let t^{proc}_j represent the time required to process packets accumulated on a controller placed in node $j \in C$ when the information of switches is migrated at the controller. Let ρ^{delay} and ρ^{capa} denote the threshold for delay limit and the threshold for capacity limit of controllers, respectively.

The notations of decision variables used in this paper are summarized in Table 2. The main decision variables are as follows. Let $x_{i,j}$ denote a decision variable which represents the ratio of the amount of load assigned from switch $i \in S$ to a controller placed in node $j \in C$ to the amount of load generated on switch $i \in S$. Let $n_{i,j}$ denote a binary decision variable which is set to one if a controller placed in node $j \in C$ is connected to switch $i \in S$, and zero otherwise. Let c_j denote a binary decision variable which is set to one if a controller is placed in node $j \in C$, and zero otherwise. Let $h_{i,j,k}$ denote a binary decision variable which is set to one if the information of switch $i \in S$ is migrated from a controller placed in node $k \in C$ to a controller placed in node $j \in C$ during the reassignment operation, and zero otherwise. Let $t_{i,j,k}$ denote a variable indicating that switch $i \in S$ is newly connected to a controller placed in node $j \in C$ and a controller placed in node $k \in C$ has information about switch $i \in S$. Let $t^{\mathrm{prop}}_{i,j}$ denote a variable indicating propagation delay between switch $i \in S$ and controller placed in node $j \in C$. Let t^{sojo}_j denote a variable indicating load-dependent sojourn time at controller placed in node $j \in C$.

Table 2 Notations of decision variables used in this paper.

Decision variable	Definition
$x_{i,j}$	Real. Ratio of the amount of load assigned from switch $i \in S$ to controller placed in node $j \in C$ to the amount of load generated on switch $i \in S$.
$n_{i,j}$	Binary. Set to one if controller placed in node $j \in C$ is connected to switch $i \in S$, and zero otherwise.
c_j	Binary. Set to one if controller is placed in node $j \in C$, and zero otherwise.
$h_{i,j,k}$	Binary. Set to one if the information of switch $i \in S$ is migrated from controller placed in node $k \in C$ to controller placed in node $j \in C$ during the reassignment operation, and zero otherwise.
$g_{j,k}$	Binary. Set to one if the information of switches is migrated between controllers placed in nodes $j, k \in C$ during the reassignment operation, and zero otherwise.
f_k	Binary. Set to one if the information of switches is migrated from controller placed in node $k \in C$ during the reassignment operation, and zero otherwise.
$y_{i,j,k}$	Binary. Set to one if $t_{i,j,k}$ is greater than zero, and zero otherwise.
$z_{i,j}$	Binary. Set to one if $y_{i,j,k}$ is at least one for all k in $i \in S$, $j \in C$, and zero otherwise.
$t_{i,j,k}$	A variable indicating that switch i is newly connected to controller placed in node $j \in C$ and controller placed in node $k \in C$ has information about switch $i \in S$.
$t^{\mathrm{prop}}_{i,j}$	A variable indicating propagation delay between switch $i \in S$ and controller placed in node $j \in C$.
t^{sojo}_j	A variable indicating load dependent sojourn time at controller placed in node $j \in C$.
ρ^{usable}	A variable indicating maximum usable processing capacity of controllers.

4.3 Assumptions about Migration Blackout Time

The migration blackout time consists of two components. The first is the time it takes to process the packets that have accumulated in the controller. The second is the time it takes to migrate the information of switches between controllers. In the proposed model, we assume that the time taken to process packets accumulated on the controller is the sum of the times that packets arriving at the controller are waiting to be processed and the time taken to be processed at the time of the migration. The sum of the time corresponds to the load-dependent sojourn time at the controller. The load-dependent sojourn time is calculated from the packet arrival rate and the packet processing rate of the controllers at the time before the reassignment operation. We also assume that the time taken to migrate the information of the switches between controllers is the propagation delay between the nodes where the controllers are placed. We set $t^{\mathrm{mig}}_{j,k}$ as the migration blackout time for sending the information of switches from that placed in node $j \in C$ to the controller placed in node $k \in C$. $t^{\mathrm{mig}}_{j,k}$ is described as:

$$t^{\mathrm{mig}}_{k,j} = t^{\mathrm{proc}}_k + t^{\mathrm{con}}_{k,j}, \forall k, j \in C. \tag{1}$$

4.4 Network Environment

In the proposed model, we assume that each node in the network has a single switch and is connected to another node by a physical link. Each node has the computational resources to host one controller. The maximum number of controllers that can be placed on each node is one. Figure 5 shows the relationship between switches, controllers, and nodes in the physical network. In this example, six nodes are connected by physical links. Controllers are deployed on nodes 2 and 6.

4.5 Delay Requirement between Switches and Controllers

We consider a constraint that the sum of the propagation

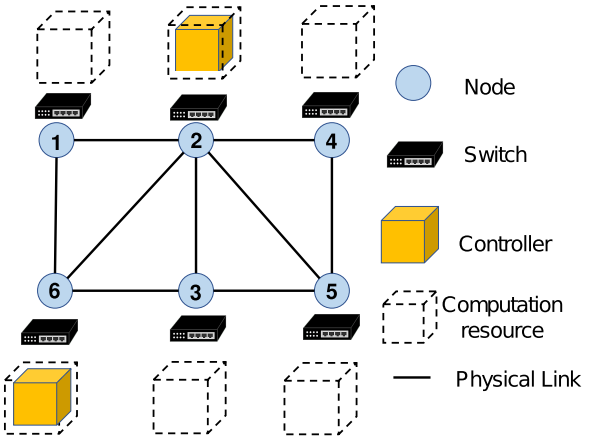


Fig. 5 Physical network.

delay between switch $i \in S$ and a controller placed in node $j \in C$ and the sojourn time at the controller is less than a specified value. The sojourn time at the controller is expressed by queueing theory based on the Kleinrock independence approximation [18]. The constraint equations are as follows:

$$2 \cdot t_{i,j}^{\text{prop}} + t_j^{\text{sojo}} \leq t^{\text{lim}}, \forall i \in S, j \in C, \tag{2a}$$

$$t_{i,j}^{\text{prop}} = n_{i,j} \cdot t_{i,j}^{\text{short}}, \forall i \in S, \tag{2b}$$

$$t_j^{\text{sojo}} = \frac{1}{\zeta_j - \alpha_j}, \tag{2c}$$

$$\alpha_j = \sum_{i \in S} d_i \cdot x_{i,j}, \forall j \in C. \tag{2d}$$

4.6 Problem Formulation

We formulate the proposed model in consideration of the four cases of reallocation. The objective function and constraints of the proposed model are described as:

$$\min \quad \sum_{k \in C} t_k^{\text{mig}} \cdot \tilde{c}_k \tag{3a}$$

$$\text{s.t.} \quad 2 \cdot t_{i,j}^{\text{prop}} + t_j^{\text{sojo}} \leq \rho^{\text{delay}} \cdot t^{\text{lim}}, \forall i \in S, j \in C, \tag{3b}$$

$$\sum_{k \in C} y_{i,j,k} \cdot h_{i,j,k} \cdot \tilde{c}_k \geq M \cdot (z_{i,j} - 1) + 1,$$

$$\forall i \in S, j \in C, \tag{3c}$$

$$t_{i,j,k} > -M \cdot (1 - y_{i,j,k}), \forall i \in S, j, k \in C, \tag{3d}$$

$$-t_{i,j,k} > (\epsilon - M) \cdot y_{i,j,k} - \epsilon, \forall i \in S, j, k \in C, \tag{3e}$$

$$h_{i,j,k} \leq g_{j,k}, \forall i \in S, j, k \in C, \tag{3f}$$

$$g_{j,k} \leq f_k, \forall j, k \in C, \tag{3g}$$

$$x_{i,j} \leq n_{i,j}, \forall i \in S, j \in C, \tag{3h}$$

$$n_{i,j} \cdot \epsilon \leq x_{i,j}. \forall i \in S, j \in C, \tag{3i}$$

$$x_{i,j} \leq c_j, \forall i \in S, j \in C, \tag{3j}$$

$$c_j \cdot \epsilon \leq x_{i,j}, \forall i \in S, j \in C, \tag{3k}$$

$$t_k^{\text{mig}} = f_k \cdot t_k^{\text{proc}} + \sum_{j \in C} g_{j,k} \cdot t_{j,k}^{\text{con}}, \forall k \in C, \tag{3l}$$

$$t_k^{\text{proc}} = \frac{1}{\zeta_j - \sum_{i \in S} d_i \cdot \tilde{x}_{i,j}}, \forall k \in C, \tag{3m}$$

$$t_{i,j,k} = (n_{i,j} - \tilde{n}_{i,j}) \cdot \tilde{n}_{i,k}, \forall i \in S, j, k \in C, \tag{3n}$$

$$t_j^{\text{sojo}} = \frac{1}{\zeta_j - \alpha_j}, \forall j \in C, \tag{3o}$$

$$t_{i,j}^{\text{prop}} = n_{i,j} \cdot t_{i,j}^{\text{short}}, \forall i \in S, j \in C, \tag{3p}$$

$$\alpha_j = \sum_{i \in S} d_i \cdot x_{i,j}, \forall j \in C, \tag{3q}$$

$$\sum_{j \in C} x_{i,j} = 1, \forall i \in S, \tag{3r}$$

$$\alpha_j \leq \rho^{\text{capa}} \cdot \zeta_j, \forall j \in C, \tag{3s}$$

$$\sum_{j \in C} c_j \leq c^{\max}, \tag{3t}$$

$$n_{i,j} \in \{0,1\}, \forall i \in S, j \in C, \tag{3u}$$

$$c_j \in \{0,1\}, \forall j \in C, \tag{3v}$$

$$h_{i,j,k} \in \{0,1\}, \forall i \in S, j, k \in C, \tag{3w}$$

$$g_{j,k} \in \{0,1\}, \forall k, j \in C, \tag{3x}$$

$$f_k \in \{0,1\}, \forall k \in C, \tag{3y}$$

$$z_{i,j} \in \{0,1\}, \forall i \in S, j \in C, \tag{3z}$$

$$y_{i,j,k} \in \{0,1\}, \forall i \in S, j, k \in C. \tag{3aa}$$

$$0 \leq x_{i,j} \leq 1, \forall i \in S, j \in C, \tag{3ab}$$

Equation (3a) shows the objective function, which minimizes the migration blackout time incurred when determining the controller assignment and placement. Equation (3b) ensures that the delay between switch $i \in S$ and a controller placed in node $j \in C$ is less than or equal to the delay limit. Equation (3c) ensures that, if switch $i \in S$ is assigned to a controller placed in node $j \in C$, the information of switch i is migrated from any controllers that have the information. Equations (3d)–(3e) ensure that (3c) holds when $t_{i,j,k}$ is less than or equal to zero. Equations (3f)–(3k) define relationships between the decision variables. ϵ is a relatively small positive constant due to a management purpose, where $0 < \epsilon \ll 1$. Equation (3l) represents the migration blackout time that occurs when migrating information of switches from a controller placed in node $k \in C$. Equation (3m) represents the time taken to process the packets queued at a controller placed in node $k \in C$ during the reassignment operation. Equation (3n) represents that $t_{i,j,k}$ becomes one when switch $i \in S$ newly connects to controller that is placed in node $j \in C$ and controller k has the information of switch $i \in S$. Equation (3o) represents the sojourn time at a controller placed in node $j \in C$. Equation (3p) represents the propagation delay between switch $i \in S$ and a controller placed in node $j \in C$. Equation (3q) represents the packet arrival rate at a controller placed in node j. Equation (3r) ensures that that decision variable $x_{i,j}$ adds up to one for all controllers that manage switch $i \in S$. Equation (3s) ensures that the arrival rate at a controller placed in node $j \in C$ is less than the product of ρ^{capa} and packet processing rate at controller $j \in C$. Equation (3t) ensures that the num-

ber of controllers placed in the network is limited to $c^{\max}$. Equations (3u)–(3ab) define the decision variable.

4.7 Equation Transformation

Since (3b) is non-linear, (3b) is transformed to a second-order conic form [13], [19].

The packet processing rate is larger than the packet arrival rate from (3s). Let p_j be defined as $\zeta_j - \alpha_j$. By multiplying both sides of (3b) by p_j, (3b) is transformed as:

$$p_j \left(\rho^{\text{delay}} \cdot t^{\text{lim}} - 2 \cdot t^{\text{prop}}_{i,j}\right) \geq 1, \forall i \in S, j \in C, \tag{4a}$$

$$p_j = \zeta_j - \alpha_j, \forall j \in C. \tag{4b}$$

Let $q_{i,j}$ be defined as $\rho^{\text{delay}} \cdot t^{\text{lim}} - 2 \cdot t^{\text{prop}}_{i,j}$. Equation (4a) can be transformed as:

$$p_j \cdot q_{i,j} \geq 1, \forall i \in S, j \in C, \tag{5a}$$

$$q_{i,j} = \rho^{\text{delay}} \cdot t^{\text{lim}} - 2 \cdot t^{\text{prop}}_{i,j}, \forall i \in S, j \in C. \tag{5b}$$

Let $r_{i,j}$ be defined as $p_j + q_{i,j}$. Equation (5a) is transformed as:

$$(r_{i,j})^2 - (p_j)^2 - (q_{i,j})^2 \geq 2, \forall i \in S, j \in C, \tag{6a}$$

$$r_{i,j} = p_j + q_{i,j}, \forall i \in S, j \in C. \tag{6b}$$

Next, since (3c) is non-linear, we transform (3c) to a liner form. Since $h_{i,j,k}$ and $t_{i,j,k}$ are binary decision variables, (3c) is transformed as follows.

$$\sum_{k \in C} v_{i,j,k} \cdot \tilde{c}_k \geq M(z_{i,j} - 1) + 1, \forall i \in S, j \in C, \tag{7a}$$

$$v_{i,j,k} \geq y_{i,j,k} + h_{k,i,j} - 1, \forall i \in S, j, k \in C, \tag{7b}$$

$$v_{i,j,k} \leq y_{i,j,k}, \forall i \in S, j, k \in C, \tag{7c}$$

$$v_{i,j,k} \leq h_{i,j,k}, \forall i \in S, j, k \in C, \tag{7d}$$

$$v_{i,j,k} \in \{0, 1\}, \forall i \in S, j, k \in C. \tag{7e}$$

To sum up, the proposed model can be treated as an MISOCP problem as:

$$\min \sum_{k \in C} t^{\text{mig}}_k \cdot \tilde{c}_k, \tag{8a}$$

s.t. (3d)–(3aa), (4b), (5b)–(7e).

5. Baseline Model

In order to evaluate the performance of our proposed model, we compare the proposed model to the model which does not consider the platform of load distribution while determining the controller placement and assignment. We introduce the model which is used in the previous research [8] as the baseline model. In the baseline model, a switch can only be connected to a single controller and the migration blackout always occurs when the controller assignment and placement are changed.

The objective function and constraints of the baseline model are described as:

$$\min \sum_{k \in C} t^{\text{mig}}_k \cdot \tilde{c}_k \tag{9a}$$

s.t. (3b)–(3g), (3l), (3n)–(3o), (3s)–(3aa),

$$n_{i,j} \leq c_j, \forall i, \in S, j \in C, \tag{9b}$$

$$\alpha_j = \sum_{i \in S} d_i \cdot n_{i,j}, \forall i \in S, j \in C, \tag{9c}$$

$$\sum_{j \in C} n_{i,j} = 1, \forall i \in S, \tag{9d}$$

$$c_j \leq \sum_{i \in S} n_{i,j}, \forall j \in C, \tag{9e}$$

$$t^{\text{proc}}_k = \frac{1}{\zeta_j - \sum_{i \in S} d_i \cdot \tilde{n}_{i,j}}, \forall k \in C. \tag{9f}$$

Equation (9b) ensures that the packets arriving at switch $i \in S$ are processed by a controller placed in node $j \in C$ in the network. Equation (9c) shows that the packet arrival rate is expressed by the total packet-in messages arriving at a controller placed in node $j \in C$. Equation (9d) guarantees that each switch is assigned to one controller. Equation (9e) ensures that no controller is placed in a node where the connections between switches and the controller are not established. Equation (9f) represents the time taken to process the packets queued at a controller placed in node $k \in C$ during the reassignment operation in the baseline model.

6. Numerical Results

6.1 Simulation Environment

In the initial placement and assignment in the evaluation of the proposed model, a switch is assumed to be connected to all placed controllers. We make this assumption in consideration of the fact that migration blackout can be avoided when connections are established between a switch and all controllers by the load-balancing platform.

We use the NSF network, which is shown in Fig. 6, in this evaluation [20]. The distances between nodes are determined based on the previous work [12], [13]. The numbers on the links represent the propagation delay of the

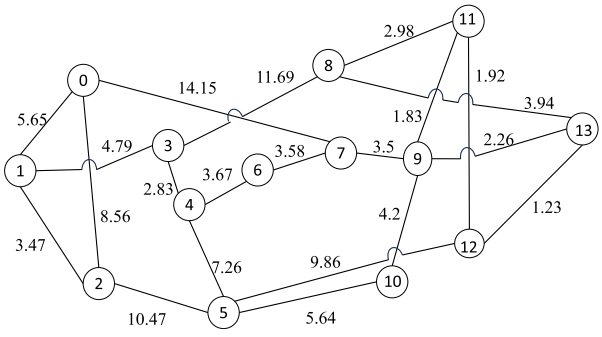


Fig. 6 NSF network. Each number shown between nodes represents propagation delay of a link [ms].

links, which are calculated by dividing the distances between the nodes by the speed of light in an optical fiber.

In the performance evaluation, we use CPLEX [21] and Python to solve the MISOCP problem. CPLEX is a tool to solve integer programming problems. We use a Dell PowerEdge R330 server with Ubuntu 18.04.1 LTS. The server equips Intel(R) Xeon(R) CPU E3-1270 v6 with 3.80 GHz. The numbers of cores and threads in the CPU are 4 and 8, respectively. The memory capacity is 64 GB.

6.2 Initial Placement and Assignment

This section shows how to determine the initial placement of controllers and assignment between switches and controllers. Switches are connected to placed controllers that can be connected to them. The number of controllers deployed in the network is limited to a specific value.

The optimization problem to determine the initial placement and initial assignment for the proposed model is as follows.

$$\max \sum_{i\in S, j\in C} n_{i,j} \tag{10a}$$

$$\text{s.t.} \quad 2\cdot t^{\text{prop}}_{i,j} + t^{\text{sojo}}_{j} \le \rho^{\text{delay}}\cdot t^{\text{lim}}, \forall i \in S, j \in C, \tag{10b}$$

$$t^{\text{sojo}}_{j} = \frac{1}{\zeta_j - \alpha_j}, \forall j \in C, \tag{10c}$$

$$\alpha_j \le \rho^{\text{capa}}\cdot \zeta_j, \forall j \in C, \tag{10d}$$

$$\alpha_j = \sum_{i\in S} d_i \cdot x_{i,j}, \forall j \in C, \tag{10e}$$

$$\sum_{j\in C} x_{i,j} = 1, \forall i \in S, \tag{10f}$$

$$x_{i,j} \le n_{i,j}, \forall i \in S, j \in C, \tag{10g}$$

$$n_{i,j} \le c_j, \forall i \in S, j \in C, \tag{10h}$$

$$c_j \cdot \epsilon \le x_{i,j}, \forall i \in S, j \in C, \tag{10i}$$

$$\sum_{j\in C} c_j \le c_{\text{initial}}, \tag{10j}$$

$$0 \le x_{i,j} \le 1, \forall i \in S, j \in C, \tag{10k}$$

$$n_{i,j} \in \{0,1\}, \forall i \in S, j \in C, \tag{10l}$$

$$c_j \in \{0,1\}, \forall j \in C. \tag{10m}$$

The objective function is to maximize the sum of binary variables representing the connections between switch $i \in S$ and controller placed in node $j \in C$. In addition, capacity and delay requirements are considered so that decision variables $x_{i,j}$ and c_j can be adopted as $\tilde{x}_{i,j}, \tilde{c}_j$ in the proposed model. The upper limit of the number of controllers is set to c_{initial}; the number of controllers is constrained to be less than or equal to c_{initial}.

The optimization problem to determine the initial placement and assignment for the baseline model is as follows.

$$\max \sum_{i\in S, j\in C} n_{i,j} \tag{11a}$$

$$\text{s.t.} \quad \text{(10b)–(10d), (10h), (10l)–(10m),}$$

$$\sum_{j\in C} c_j = c_{\text{initial}}, \tag{11b}$$

$$c_j \cdot \epsilon \le n_{i,j}, \forall i, \in S, j \in C, \tag{11c}$$

$$\alpha_j = \sum_{i\in S} d_i \cdot n_{i,j}, \forall i \in S, j \in C, \tag{11d}$$

$$\sum_{j\in C} n_{i,j} = 1, \forall i \in S. \tag{11e}$$

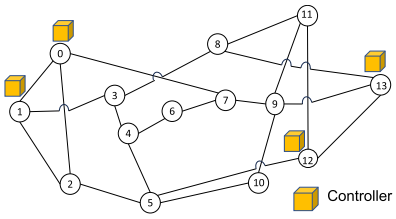
6.3 Demonstration of Proposed and Baseline Models

We show the demonstration of the proposed and baseline models for the case where the migration blackout cannot be avoided.

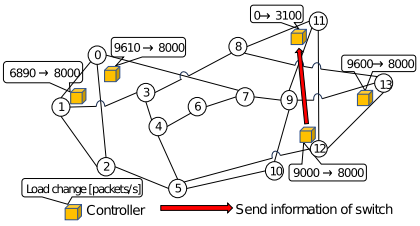
In the demonstration, t^{lim}, which denotes the upper bound of the delay between switches and controllers, is set to 100 [ms]. ζ_j, which denotes the processing rate of the controller, is set to 10×10^3 [packets/s]. ρ^{delay}, which denotes the threshold for delay limit, is set to 0.8. ρ^{capa}, which denotes the threshold for capacity limit of controllers, is set to 0.8. The threshold of the delay between switches and controllers is $\rho^{\text{delay}}\cdot t^{\text{lim}} = 80$ [ms]. The threshold of the controller's processing capacity is $\rho^{\text{capa}}\cdot\zeta_j = 8000$ [packets/s].

Figure 7(a) shows the initial controller placement for the proposed and baseline models. The controller assignment is determined based on the initial controller placement. The value of d_i needs to satisfy constraints (10b)–(10e) in the case of the proposed model and (11d) in the case of the baseline model. We set d_i so that the load that exceeds the capacity threshold is assigned to some of the controllers and the load close to the threshold is assigned to the other controllers. In the initial placement and assignment for the proposed model, the load that exceeds the capacity threshold is assigned to the controllers placed in nodes 0, 12, and 13. In the baseline model, the load that exceeds the threshold of capacity is assigned to the controllers placed in nodes 0, 1, and 13.

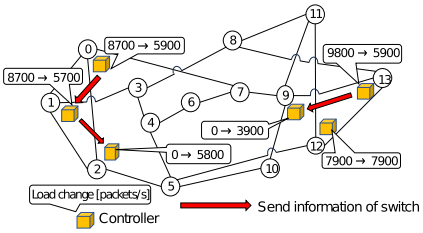
Figures 7(b) and 7(c) show how the information of switches is migrated between controllers as a result of using the proposed and baseline models, respectively. Table 3 summarizes the demonstration result. The result shows that, in the proposed model, a new controller is placed in node 11 and the information of switches is migrated to the newly placed controller. The reasons for this are as follows. When a load exceeding the capacity threshold is assigned to one controller, the load needs to be reassigned to other controllers. When a load close to the capacity threshold is reassigned to the controllers, it is not likely that the load is handled only by changing the ratio of the amount of load assigned from switches to controllers; in this case, a new controller needs to be placed to migrate the information of switch between controllers. The result also shows that the migration blackout time in the proposed model is shorter than that in the baseline model in Table 3. This is because, in the proposed model, each controller has the information of all switches in the initial placement, and the information of switches can



(a) Initial controller placement for proposed and baseline models.



(b) Load changes by proposed model.



(c) Load changes by baseline model.

Fig. 7 Initial controller placement and changes of load assigned to controllers determined by proposed and baseline models.

Table 3 Demonstration result.

Model	Initial controller placement	Determined controller placement	Migration blackout time [ms]
Proposed model	Nodes 0, 1, 12, 13	Nodes 0, 1, 11, 12, 13	2.92
Baseline model		Nodes 0, 1, 2, 9, 12, 13	17.92

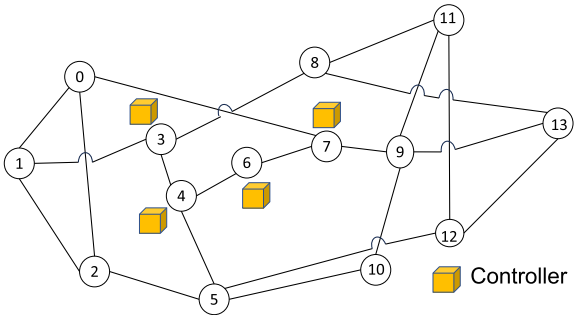


Fig. 8 Initial controller placement for proposed model and baseline model in second situation.

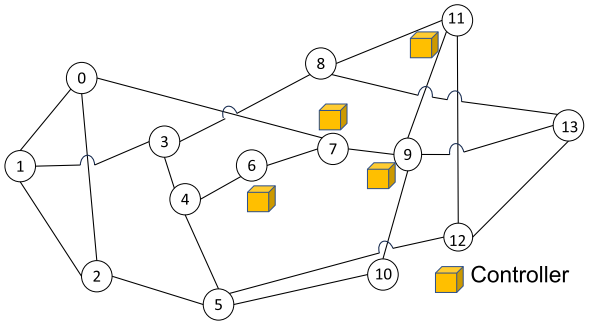


Fig. 9 Initial controller placement for proposed model and baseline model in third situation.

be migrated from any controller that is placed in the initial assignment. In the baseline model, only one controller has information of a particular switch, so the switch information can only be migrated from the controller.

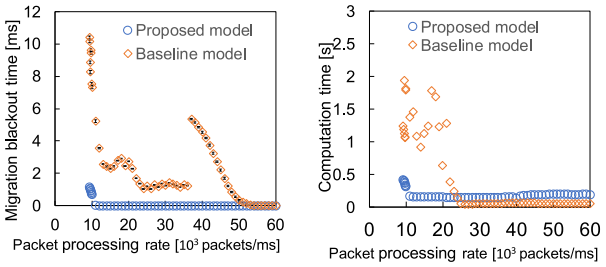
6.4 Static Scenario

This section evaluates the performance of the proposed model in the case where a load of each switch at a certain time is given. The initial placement of controllers and the initial assignment between switches and controllers are given in three situations. In the first situation, the initial placement and assignment are the same as those in Sect. 6.3. In the second situation, the initial placement of controllers is determined so that the maximum distance from each node to the furthest controller is small. The initial assignment is determined based on the initial placement. Figure 8 shows the initial placement of controllers in the second situation. In the third situation, controllers are placed at nodes with high closeness centrality in the initial placement. The closeness centrality is a type of network centrality [22]. The closeness centrality is determined by the average length of the shortest path from each node to the other. The initial assignment is determined based on the initial placement. Figure 9 shows the initial placement of controllers in the third situation.

In the evaluation, the upper bound of the delay between switches and controllers, t^{lim}, is set to 100 [ms], which is sufficiently large compared to the propagation delay in Fig. 6 to make each switch connectable to all controllers. The amount of packets arriving at each switch, d_i, is randomly set in the range of 1×10^3 to 4×10^3 [packets/s]. d_i is set at random since packets do not arrive equally at each switch in the actual network. The threshold for delay limit, ρ^{delay}, and the threshold for capacity limit of controllers, ρ^{capa}, are set to 0.8. The processing rate of the controller, ζ_j, is changed from 9.3×10^3 to 60×10^3 [packets/s]. The number of simulation trials is set to 10000. The maximum amount of load assigned to the controller is denoted by ρ^{limit}. ρ^{limit} represents the maximum usable processing capacity of each controller. ρ^{limit} is set to 0.95. If the amount of load assigned to a controller becomes larger than the product of ρ^{limit} and ζ_j, the proposed and baseline models newly determine the controller assignment and placement. By setting ρ^{limit} appropriately, the network operator can avoid large sojourn time at the controller which can be caused by assigning load close to its capacity.

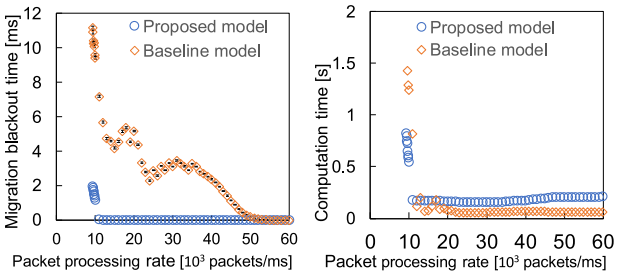
6.4.1 Result in First Situation

In this section, the migration blackout time and the computation time are evaluated in the first situation.



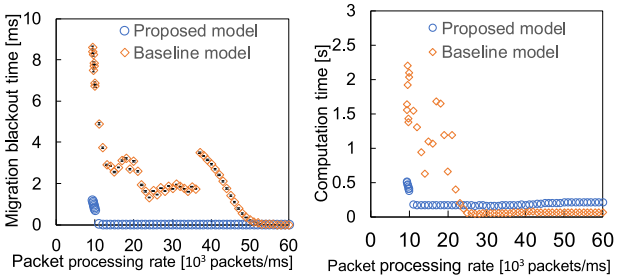
(a) Migration blackout time. (b) Computation time.

Fig. 10 Processing rate dependency of migration blackout time and computation time in proposed models in first situation.



(a) Migration blackout time. (b) Computation time.

Fig. 11 Processing rate dependency of migration blackout time and computation time in proposed models in second situation.



(a) Migration blackout time. (b) Computation time.

Fig. 12 Processing rate dependency of migration blackout time and computation time in proposed models in third situation.

Figure 10(a) shows the processing rate dependency of the migration blackout time in the proposed and baseline models in the first situation. Error bars indicate the width of the confidence interval. The result shows that the proposed model requires less migration blackout time than the baseline model. This is because the proposed model can determine the amount of load allocated to a controller while avoiding the migration blackout. In the proposed model, when the packet processing rate becomes smaller, the migration blackout time becomes larger. This indicates that, when the packet processing rate decreases, a new controller is placed and switches are connected to the controller, which results in the migration of switch information between controllers. In the baseline model, a migration blackout always occurs when the amount of load assigned to the controller becomes larger than ρ^{capa} and the controller load is reassigned to another controller. The result also shows that the migration blackout time tends to approach zero in the proposed and baseline models when the packet processing rate becomes sufficiently large. This is because, as the processing rate increases, the amount of load that the controller can handle increases, which decreases the frequency of the migration blackout. When $15 \times 10^3 \leq \zeta_j \leq 18 \times 10^3, 26 \times 10^3 \leq \zeta_j \leq 31 \times 10^3$, and $36 \times 10^3 \leq \zeta_j \leq 37 \times 10^3$, the migration blackout time in the baseline model increases as the packet processing rate increases. The reason for this is as follows. As the packet processing rate increases, each controller can control more switches and the number of controllers used to control switches decreases. When the number of controllers used in the initial placement decreases, the propagation delay between switches and controllers tends to be larger.

Figure 10(b) shows the processing rate dependency of the computation time in the proposed and baseline models in the first situation. The result shows that, in the proposed model, the computation time tends to decrease as the packet processing rate of the controller increases. This is because, as the processing rate increases, the amount of load that the controller can accommodate increases, which makes it easier to find a solution that avoids the migration blackout. The result also shows that, when $23 \times 10^3 \leq \zeta_j \leq 35 \times 10^3$, the computation time of the baseline model becomes smaller than that of the proposed model. This is because, as the processing rate becomes larger, a single controller can handle more load in the initial placement, which results in reducing the number of candidate controllers for migrating switch information and, as a result, the time required to search for the optimal solution decreases.

6.4.2 Results in Different Situations

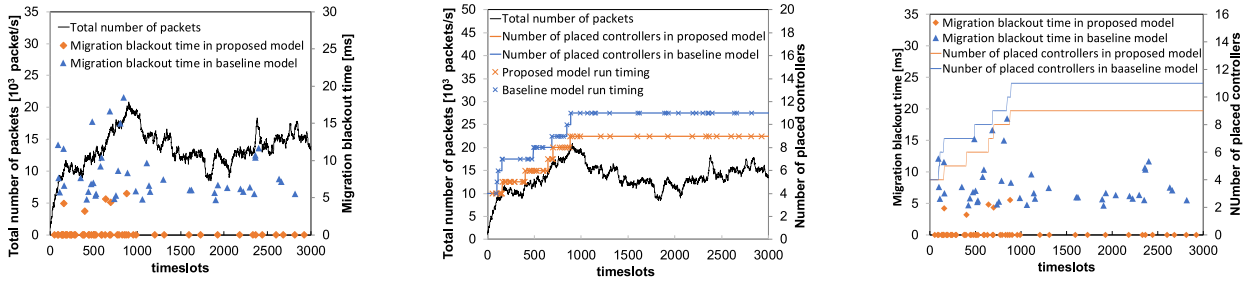
In this section, the migration blackout time and the computation time are evaluated in the second and third situations.

Figure 11(a) shows the processing rate dependency of the migration blackout time in the proposed and baseline models in the second situation. Fig. 11(b) shows the processing rate dependency of the computation time in the proposed and baseline models. The result trend in Figs. 11(a) and 11(b) is the same as that in the first situation. The result in Fig. 11(b) shows that, when $13 \times 10^3 \leq \zeta_j \leq 60 \times 10^3$, the computation time of the baseline model becomes smaller than that of the proposed model.

Figure 12(a) shows the processing rate dependency of the migration blackout time in the proposed and baseline models in the third situation. Fig. 12(b) shows the processing rate dependency of the computation time in the proposed and baseline models. The result trend in Figs. 12(a) and 12(b) is the same as that in the first situation.

6.5 Dynamic Scenario

This section evaluates the performance of both proposed and baseline models under the scenario where the load generated



(a) Total number of packets and migration blackout time.

(b) Total number of packets and number of placed controllers.

(c) Migration blackout time and number of placed controllers.

Fig. 13 Result of dynamic scenario in first situation.

on the switches varies. The initial placement and assignment of the controllers are determined in the same way as in Sect. 6.4.

In the evaluation, the upper bound of the delay between switches and controllers, t^{lim}, is set to 70 [ms]. The amount of load generated on the switches in timeslot t is determined by the sum of the amount of load in timeslot $t-1$ and the amount of variation. The amount of variation is generated by using a truncated normal distribution. The threshold for delay limit and that for capacity limit of controllers, ρ^{delay} and ρ^{capa}, are set to 0.80. The processing rate of each controller, ζ_j, is set to 3×10^3 [packets/s]. The maximum amount of load assigned to the controller, ρ^{limit}, is set to 0.90. The maximum number of controllers placed in the evaluation, $c^{\max}$, is set to 11. The maximum number of timeslots in the evaluation is 3000.

We assume the scenario that the network operator does not remove the controller that once placed in the network. The reason for this is that, if a controller that is once placed is removed and then placed again, the information of switches needs to be migrated to the newly placed controllers, which causes the migration blackout. The following equation is added to the proposed and baseline models for the simulation in the dynamic scenario.

$$\sum_{j \in C} \tilde{c}_j \leq \sum_{j \in C} c_j. \tag{12}$$

Equation (12) ensures that the number of placed controllers does not decrease.

The objective function is modified into the following Eq. (13a) in order to reduce the times that the optimization problems of the proposed and baseline models are solved. ρ^{usable} denotes a decision variable representing the threshold of the controller's capacity limit. μ is a constant that is sufficiently small compared to the first term of the objective function.

$$\min \sum_{k \in C} t_k^{\text{mig}} \cdot \tilde{c}_k + \mu \cdot \rho^{\text{usable}}, \tag{13a}$$

$$0 \leq \rho^{\text{usable}} \leq \rho^{\text{capa}}, \tag{13b}$$

$$\alpha_j \leq \rho^{\text{usable}} \cdot \zeta_j, \forall j \in C, \tag{13c}$$

Equation (13a) shows the new objective function. Equation (13b) defines ρ^{usable}. Equation (13c) ensures that the arrival rate at a controller placed in node $j \in C$ is less than the product of ρ^{usable} and packet processing rate at controller $j \in C$. By modifying the objective function and introducing these constraint equations, the amount of load allocated to the controller is equalized, which results in reduction of the number of times the model runs.

6.5.1 Result in First Situation

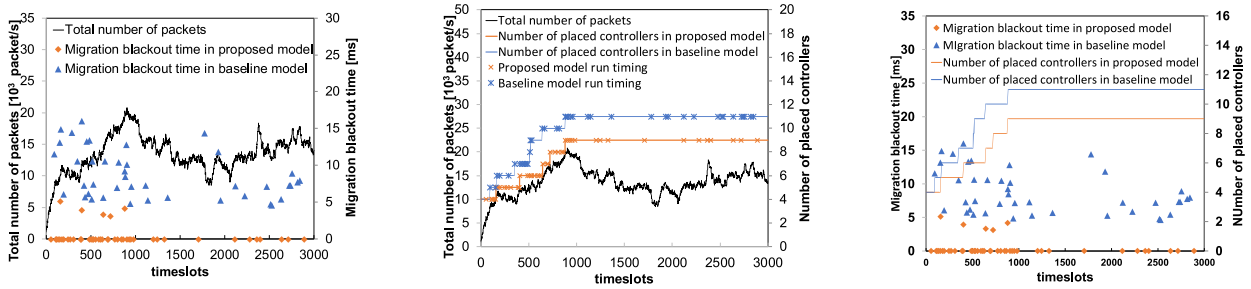
In this section, the migration blackout time and the number of placed controllers are evaluated in the first situation. Figures 13(a)–13(c) show the results of the migration blackout time and the number of controllers placed in the first situation. Figure 13(a) shows that the proposed model achieves less average migration blackout time than the baseline model. Figure 13(b) shows that the number of placed controllers in the proposed model is smaller than that in the baseline model. This is because the proposed model can use the processing capacity of the controllers more efficiently than the baseline model. Figure 13(c) shows that the migration blackout occurs when the number of controllers in the proposed model increases. This is because, when a new controller is placed in the proposed model, the switch information needs to be migrated to the controller, which results in occuring migration blackout time.

6.5.2 Results in Different Situations

In this section, the migration blackout time and the number of placed controllers are evaluated in the second and third situations.

Figures 14(a)–14(c) show the result of the migration blackout time and the number of controllers placed in the network in the proposed and baseline models. The result trend in Figs. 14(a) and 14(b) are the same as those in the first situation. Figure 14(c) shows that the speed of increase in the number of controllers in the baseline model is faster than that in the first situation.

Figures 15(a)–15(c) show the result in the proposed and baseline models in the third situation. Figure 15(a) shows that there are periods of time when the models do not run and the migration blackout time does not occur. The reason for

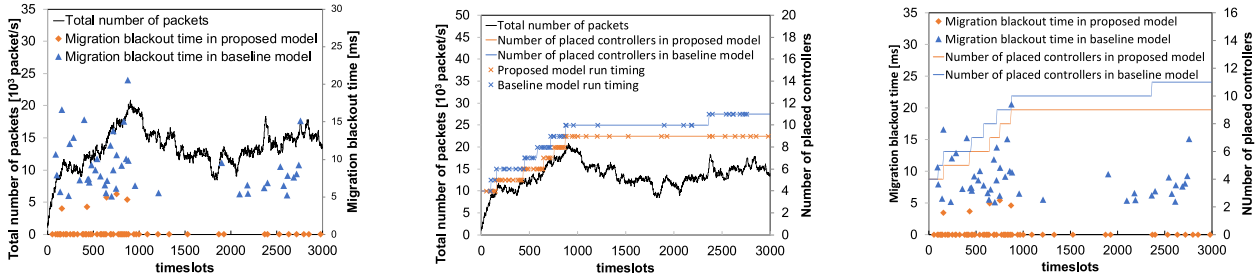


(a) Total number of packets and migration blackout time.

(b) Total number of packets and number of placed controllers.

(c) Migration blackout time and number of placed controllers.

Fig. 14 Result of dynamic scenario in second situation.

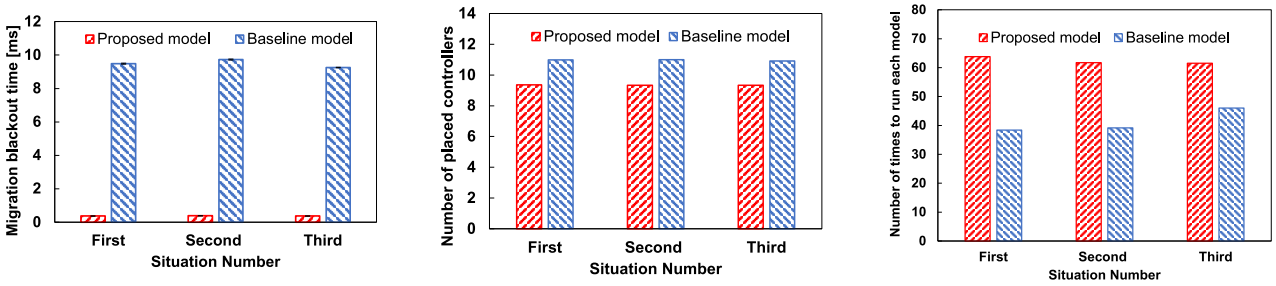


(a) Total number of packets and migration blackout time.

(b) Total number of packets and number of placed controllers.

(c) Migration blackout time and number of placed controllers.

Fig. 15 Result of dynamic scenario in third situation.



(a) Migration blackout time.

(b) Number of placed controllers.

(c) Number of times to run each model.

Fig. 16 Result of multiple scenarios in dynamic situation.

this is that the amount of packets generated on the switches in the network is less than the most recent peak value. The result also shows that the number of controllers placed in the network in the baseline model in the second situation is smaller than that in the first situation. The trend of the results in Figs. 15(a)–15(c) is the same as those in the first situation.

6.5.3 Evaluation in Multiple Scenarios

In this section, the migration blackout time, the number of controllers that are placed when the timeslot reaches 3000, and the number of times to run each model are evaluated in each of the three situations in multiple scenarios. The number of scenarios in this evaluation is 100. The settings of the constants are the same as those in Sect. 6.5.1.

Figures 16(a)–16(c) show the results of the proposed and baseline models in multiple scenarios. Error bars indicate the width of the confidence interval. Figure 16(a) shows that the proposed model newly determines network connections with smaller migration blackout time than the baseline model. This is because the proposed model is more likely to avoid the migration blackout when the controller assignment and placement are newly determined. Figure 16(b) shows that the number of placed controllers in the proposed model is smaller than that in the baseline model. Figure 16(c) shows that the number of times to run the baseline model is smaller than that of the proposed model. The reason for this is as follows. The number of placed controllers to handle the load generated by the switches in the proposed model is smaller than that in the baseline model. This means that less

controller processing capacity is actually available than the baseline model. Therefore, the amount of load allocated to each controller in the proposed model is larger than that in the baseline model, which increases the number of times to run the proposed model.

6.6 Features of Proposed Model

This section discusses the pros and cons of the proposed model.

In the proposed model, when the number of packets arriving at switches fluctuates, there are two cases to handle the situation. One case is that the existing controllers can handle the fluctuation of the number of packets, and the other case is that additional controllers need to be deployed to handle the fluctuation. In the former case, the information of switches in the network is shared by all the deployed controllers with a load-balancing platform. Therefore, the migration of switch information between controllers in not required when reassigning switches. This can avoid the migration blackout. In addition, each switch can be controlled by multiple controllers with the load-balancing platform. This allows the processing capacity of each controller to be used efficiently. The migration blackout occurs when additional controllers are deployed in the proposed model. However, the platform achieves load balancing, which reduces the number of situations where additional controllers are needed.

On the other hand, the proposed model has some cons. The proposed model assumes that once the number of controllers is increased, the number of controllers does not decrease, for the purpose of minimizing the migration blackout time. When the number of packets in the network decreases, the number of controllers does not decrease, and controllers may be over-allocated. In addition, the number of times the model runs for reallocation increases with the load-balancing platform. This increases the amount of work involved in changing the configuration of the network. We recognize these cons as future work.

7. Conclusion

This paper proposed a model that determines the controller assignment and placement while minimizing the migration blackout time with the load-balancing platform. We formulated the proposed model as a mixed-integer second-order cone programming problem. We developed a migration procedure used in the proposed model. In the procedure, each switch can be controlled by multiple controllers with a load-balancing platform. The load-balancing platform allows status messages sent from a switch to be sent to multiple controllers. This allows multiple controllers to process messages sent from the switch in a proper order and the migration blackout time can be avoided. This paper evaluated the performance of the proposed model in both static and dynamic scenarios. Numerical results showed that the migration blackout time that occurs in the proposed model is smaller than that in the baseline model. This is because, in the proposed model, the migration blackout time can be avoided if the amount of load allocated from the switch to the controllers is changed. The results also showed that the number of controllers placed in the proposed model is smaller than that in the baseline model. This is because the proposed model can distribute the load of switches to multiple controllers and can use the processing capacity of each controller more efficiently than the baseline model.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Numbers 21H03426 and 23H03382, and JST, PRESTO Grant Number JPMJPR23P4, Japan.

References

[1] J.H. Cox, J. Chung, S. Donovan, J. Ivey, R.J. Clark, G. Riley, and H.L. Owen, "Advancing software-defined networks: A survey," IEEE Access, vol.5, pp.25487–25526, Oct. 2017.
[2] N. Bizanis and F.A. Kuipers, "SDN and virtualization solutions for the internet of things: A survey," IEEE Access, vol.4, pp.5591–5606, Sept. 2016.
[3] A. Nygren, B. Pfaff, B. Lantz, B. Heller, C. Barker, C. Beckmann, et al., "OpenFlow switch specification version 1.5.1," Open Networking Foundation Technical Report, 2015.
[4] R. Sherwood, G. Gibb, K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, "FlowVisor: A network virtualization layer," OpenFlow Switch Consortium Technical Report, 2009.
[5] A. Al-Shabibi, M.D. Leenheer, M. Gerola, A. Koshibe, W. Snow, and G. Parulkar, "OpenVirtex: A network hypervisor," Open Networking Summit, 2014.
[6] A. Blenk, A. Basta, and W. Kellerer, "HyperFlex: An SDN virtualization architecture with flexible hypervisor function allocation," Proc. Int. Symp. Integr. Netw. Manag., pp.397–405, July 2015.
[7] A. Dixit, F. Hao, S. Mukherjee, T.V. Lakshman, and R.R. Kompella, "ElastiCon: An elastic distributed SDN controller," Proc. ACM/IEEE Symp. Architectures Netw. Commun. Syst., pp.17–27, Oct. 2014.
[8] Y. Xu, M. Cello, I.-C. Wang, A. Walid, G. Wilfong, C.H.-P. Wen, M. Marchese, and H.J. Chao, "Dynamic switch migration in distributed software-defined networks to achieve controller load balance," J. Sel. Areas Commun., vol.37, no.3, pp.515–529, March 2019.
[9] G. Yue, Y. Wang, and Y. Liu, "Rule placement and switch migration-based scheme for controller load balancing in SDN," Proc. IEEE Symp. Comput. Commun., pp.1–6, July 2022.
[10] Z. Min, H. Sun, S. Bao, A.S. Gokhale, and S.S. Gokhale, "A self-adaptive load balancing approach for software-defined networks in IoT," Proc. IEEE Int. Conf. Autonom. Comput. Self-Org. Syst., pp.11–20, Sept. 2021.
[11] M.F. Bari, A.R. Roy, S.R. Chowdhury, Q. Zhang, M.F. Zhani, R. Ahmed, and R. Boutaba, "Dynamic controller provisioning in software defined networks," Proc. 9th Int. Conf. Netw. Service Manag., pp.18–25, Jan. 2013.
[12] S. Kotachi, T. Sato, R. Shinkuma, and E. Oki, "Fault-tolerant controller placement model by distributing switch load among multiple controllers in software-defined network," IEICE Trans. Commun., vol.E105-B, no.5, pp.533–544, May 2022.
[13] S. Noda, T. Sato, and E. Oki, "Fault-tolerant controller placement model considering load-dependent sojourn time in software-defined network," IEEE Trans. Netw. Service Manag., vol.20, no.4, pp.4887–4908, 2023, doi: 10.1109/TNSM.2023.3284830.
[14] S. Hegde, R. Ajayghosh, S.G. Koolagudi, and S. Bhattacharya, "Dynamic controller placement in edge-core software defined networks,"

Proc. IEEE Reg. 10 Annu. Int. Conf., pp.3153–3158, Nov. 2017.
[15] N. Mouawad, R. Naja, and S. Tohme, "Optimal and dynamic SDN controller placement," Proc. Int. Conf. Comput. Appl., pp.1–9, Aug. 2018.
[16] Y. Liu, H. Gu, X. Yu, and J. Zhou, "Dynamic SDN controller placement in elastic optical datacenter networks," Proc. 2018 Asia Commun. Photon. Conf., pp.1–3, Oct. 2018.
[17] S.A. Zadeh, F. Zandi, M. Buckley, and Y. Ganjali, "Meta-migration: Reducing switch migration tail latency through competition," Proc. IFIP Netw. Conf., pp.1–9, July 2023.
[18] D.P. Bertsekas and R.G. Gallager, *Data Networks,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, USA, 1992.
[19] F. He and E. Oki, "Robust virtual network function deployment against uncertain traffic arrival rates," IEEE Conf. Netw. Softwarization, pp.339–347, July 2021.
[20] K.D. Frazer, "NSFNET: A partnership for high-speed networking final report," 1995. [Online]. Available: https://www.merit.edu/wp-content/uploads/2019/06/NSFNET-final-1.pdf [Accessed: June 13, 2022]
[21] IBM ILOG CPLEX optimization studio. (2022). IBM. Accessed: Nov. 21, 2023. [Online]. Available: https://www.ibm.com/products/ilog-cplex-optimization-studio
[22] F.A. Rodrigues, "Network centrality: An introduction," A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems, pp.177–196, Springer, 2019.

Shinji Noda is pursuing the M.E. degree at Graduate School of Informatics, Kyoto University, Kyoto, Japan. He received the B.E. degree from Kyoto University, Japan, in 2022. His research interests include optimization, software-defined network, and controller placement problem.

Takehiro Sato received his B.E., M.E., and Ph.D. in engineering from Keio University in 2010, 2011, and 2016, respectively. He is currently an associate professor in the Graduate School of Informatics at Kyoto University. His research interests include design and control methods for optical and virtualized networks.

Eiji Oki is a Professor at Kyoto University, Kyoto, Japan. He was with Nippon Telegraph and Telephone Corporation (NTT) Laboratories, Tokyo, from 1993 to 2008, and The University of Electro-Communications, Tokyo, from 2008 to 2017. From 2000 to 2001, he was a Visiting Scholar at Polytechnic University, Brooklyn, New York. His research interests include routing, switching, protocols, optimization, and traffic engineering in communication networks.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

Virtual Machine Placement Method with Compressed Sensing-Based Traffic Volume Estimation*

Kenta YUMOTO[†], **Ami YAMAMOTO**[†], ***Nonmembers***, **Takahiro MATSUDA**[†a)], ***Senior Member***, **Junichi HIGUCHI**[††], **Takeshi KODAMA**[††], **Hitoshi UENO**[††], ***and*** **Takashi SHIRAISHI**[†††], ***Nonmembers***

SUMMARY In cloud computing environments with virtual machines (VMs), we propose a VM placement (VMP) method based on traffic estimation to balance loads due to traffic volumes within physical hosts (PHs) and passing through physical network interface cards (NICs). We refer to a VM or a NIC in a cloud environment as *node*, and define a *flow* as a pair of nodes. To balance loads for both PHs and NICs, it is necessary to measure flow traffic volumes because each VM may connect to other VMs in different PHs. However, this is not a cost-effective way to measure flow traffic volumes because the number of flows increases with $O(N^2)$ for the number N of nodes. To solve this problem, we propose a VMP method using a compressed sensing (CS)-based traffic estimator. In the proposed method, the relationship between flow traffic volumes and node traffic volumes is formulated by a system of underdetermined linear equations. The flow traffic volumes are estimated with CS from the measured node traffic volumes. From the estimated flow traffic volumes, each VM is assigned to the optimal host for load balancing by solving a mixed-integer optimization problem.

key words: *virtual machine, compressed sensing, VM placement, traffic volume estimation*

1. Introduction

Cloud computing technology, which provides computer resources as a service through a network, has made great progress and has become an indispensable technology in modern life. As long as users have access to the Internet, they can use the services provided on the server and access the same data regardless of their locations or computer devices. Cloud computing provides infrastructure, platforms, and software (applications) to consumers, utilizing virtualization technology to provide storage, servers, network services, CPU, and memory [1], [2].

Virtual machines (VMs) are important components in cloud computing. In this study, for a cloud computing environment with multiple physical hosts, we consider the *VM placement (VMP) problem* to deploy VMs to physical hosts. VMP methods have been studied for different objective functions and can be classified into *resource-aware*, *power-aware*, *network-aware* (or *traffic-aware*), and *cost-aware* methods [3]. Some of the many proposed methods to solve the VMP problem [3]–[19] are reviewed in Sect. 2.

In this paper, we propose a traffic-aware VMP method to balance loads due to traffic volumes between VMs [20]. The proposed method is designed to determine the optimum mapping of VMs to physical hosts (PHs) for load balancing according to inter-VM traffic volumes. To the best of our knowledge, existing traffic-aware VMP methods obtain the optimum mapping under the assumption that the inter-VM traffic volumes can be measured. However, this approach is not cost-effective because the number of VM–VM pairs increases with $O(N^2)$ for the number N of VMs. In the proposed method, instead of measuring inter-VM traffic volumes, which we refer to as *flow traffic volumes*, they are estimated from traffic volumes transmitted and received by VMs, which we refer to as *node traffic volumes*. The relationship between node traffic volumes and flow traffic volumes can be formulated as a system of underdetermined linear equations, which means that there are an infinite number of solutions for flow traffic volumes [21]. Therefore, the proposed method estimates the flow traffic volumes with compressed sensing (CS) [21], [22], which can solve the underdetermined linear inverse problem if the flow traffic volumes can be regarded as a sparse vector. From the estimated flow traffic volumes, each VM is assigned to a PH by solving a mixed integer optimization problem.

The contributions of this study are as follows.

- We propose using a CS-based estimator for flow traffic volume in the VMP problem. While CS-based traffic estimation methods have been extensively studied in the literature [24], [25], as well as the traffic-aware VMP methods, there have been no methods combining a traffic estimator and a VMP method. In addition to the ability to solve the underdetermined linear inverse problem, the estimator is well-suited to the VMP problem because smaller flow traffic volumes are estimated as zeros. The estimator is well-suited to the VMP problem because it can easily identify flows with larger traffic volumes, which significantly affect the congestion in the network.
- To balance traffic loads in PHs and physical network interface cards (NICs), we formulate a mixed-integer nonlinear optimization problem and apply estimated flow traffic volumes to it. We further extend the optimization problem to apply the VMP problem for VMs

Manuscript received March 11, 2024.
Manuscript revised June 7, 2024.
Manuscript publicized July 18, 2024.

[†]Graudate School of Systems Design, Tokyo Metropolitan University, Hino-shi, 191-0065 Japan.
[††]Fsas Technologies Inc., Kawasaki-shi, 211-0012 Japan.
[†††]Fujitsu Ltd., Kawasaki-shi, 211-8588 Japan.

*This paper is presented in part at the International Conference on Consumer Electronics - Taiwan 2023 (ICCE-TW 2023) [20].

a) E-mail: takahiro.m@tmu.ac.jp

DOI: 10.23919/transcom.2024EBT0001

Table 1 Notation.

N	number of virtual machines (VMs)
M	number of physical hosts (PHs)
N_{G}	number of groups
T	measurement period
$\mathcal{V}^{(\mathrm{VM})}$	set of VMs
$\mathcal{V}^{(\mathrm{NIC})}$	set of NICs
$\mathcal{V} = \mathcal{V}^{(\mathrm{VM})} \cup \mathcal{V}^{(\mathrm{NIC})}$	set of nodes (VMs and NICs)
$\mathcal{H}_m$	set of VMs assigned to the m-th PH
$\mathcal{V}_i$ $(i = 1, 2, \ldots, N_{\mathrm{G}})$	set of VMs in the i-th group
$\boldsymbol{x}^{(\mathrm{tx})}$	vector whose elements are traffic volumes sent from VMs
$\boldsymbol{x}^{(\mathrm{rx})}$	vector whose elements are traffic volumes received in VMs
$\boldsymbol{y}^{(\mathrm{in})}$	vector whose elements are traffic volumes injected into PHs
$\boldsymbol{y}^{(\mathrm{out})}$	vector whose elements are traffic volumes sent from PHs
$\boldsymbol{z}(t)$	node traffic vector at time t.
$\boldsymbol{u}(t)$	flow traffic vector at time t
$\hat{\mathcal{U}}$	set of estimated flow traffic volumes
$\boldsymbol{p} \in \{0, 1\}^{MN \times 1}$	vector to represent a VM placement
$\boldsymbol{q} = (q_1\ q_2\ \boldsymbol{p}^\top)^\top$	optimization variable for the problems $\mathcal{P}_1$ and $\mathcal{P}_2$
$\mathcal{V}_{\mathrm{fix}}$	set of light-traffic VMs

with larger flow traffic volumes.

In our conference paper [20], we presented the basic concept of the proposed VMP method. The proposed method consists of the traffic estimator and the assignment of VMs to PHs. In this study, we extend the traffic estimator by adding additional constraints to the optimization problem of the flow traffic estimator. We also extend the VM assignment algorithm to the max-min optimization problem from the minimization problem considered in [20].

The structure of this paper is as follows. In Sect. 2, we review the related work in this study. In Sect. 3, we describe the proposed VM placement method. We first explain the system model and overview of the proposed method, and then explain the CS-based flow traffic estimator and the mixed-integer optimization problem for the VM placement problem. In Sect. 4, we evaluate the proposed method with simulation experiments. Finally, in Sect. 5, we summarize this study.

Notation: In this study, lowercase and uppercase bold characters denote vectors and matrices, respectively. For $k = 1, 2$, the ℓ_k-norm $\|\boldsymbol{s}\|_k$ of $\boldsymbol{s} = (s_1\ s_2\ \cdots\ s_N)^\top$ is defined as

$$\|\boldsymbol{s}\|_k = \left(\sum_{n=1}^{N} |s_n|^k \right)^{1/k}.$$

The notation used in this study is summarized in Table 1.

2. Related Work

As described in Sect. 1, VMP methods are classified as resource-aware, power-aware, network-aware (or traffic-aware), and cost-aware methods. In this section, we review the traffic-aware methods [5], [7]–[10], [13].

Fang et al. [5] proposed *VMPlanner*, a VMP method to reduce the power consumption due to network elements, such as switches and links. In VMPlanner, VMP is optimized to turn off as many unneeded network elements as possible. Li et al. [7] defined the *PM-cost* and *N-cost*, where the former is proportional to the number of PHs and the latter is mainly determined by the traffic volumes between VMs. The VMP problem is formulated by accounting for the PM-cost and N-cost. Tang and Pan [8] proposed a VMP method to improve energy efficiency in data center networks. The VMP problem is formulated with the CPU usage, memory usage, and the amount of data transferred in the networks, and solved with a hybrid genetic algorithm. Ilkhechi et al. [9] proposed a VMP method that accounts for traffic volumes and introduced a new metric, *satisfaction*, which reflects the performance of a VM when it is placed on a particular PH.

Some network-aware methods formulated the VMP problem as a *multi-objective* optimization problem. Zheng et al. [10] formulate the VMP problem as a multi-objective optimization problem to minimize resource waste and power consumption with CPU usage, memory usage, datastore I/O, and network usage. Although traffic volumes between VMs are not considered in this method, an extended model with the traffic volumes is discussed. Qin et al. [13] formulated the VMP problem as a multi-objective optimization problem to maximize communication revenue and minimize PH power consumption.

In the existing network-aware VMP methods, it is assumed that flow traffic volumes are measurable. However, it is not cost-effective to measure traffic volumes because the number of flows increases with $O(N^2)$ as described in Sect. 1. Our proposed method estimates flow traffic volumes from node traffic volumes, which can be measured with $O(N)$. Some of the existing methods consider the VMP problem over rich-connected network topologies, such as Fat-Tree, VL2, and VL2N-Tree [3], [5]. In this study, we do not consider the VMP problem such a topology but rather a load-balancing problem among PHs over a simple network topology with one switch. However, the proposed VMP method can be applied to more complicated topologies because it provides a general framework to estimate flow traffic volumes.

The objective of this study is to propose a method for estimating flow traffic volumes from node traffic volumes in VMP problems. It is not our intention to overcome the existing VMP methods. By applying the proposed flow traffic estimator to the existing VMP methods using flow traffic volumes, we can construct a cost-effective VMP method. Therefore, in Sect. 4, we do not compare the performance of the proposed method with that of other methods. Instead,

we evaluate the performance by comparing it with the ideal case, where flow traffic volumes are perfectly estimated.

It is worth noting that the proposed flow traffic estimator is based on an idea similar to *traffic matrix (TM) estimation*, which has been extensively studied thus far [23]–[26]. A TM is a non-negative matrix in which each element describes the traffic volume between a source node and a destination node. The TM estimation problem can be formulated as a linear inverse problem to estimate the TM from measured link traffic volumes. Similarly, in the traffic estimator in this study, flow traffic volumes are estimated from node traffic volumes. In [24], [25], under the assumption that the sequences of traffic volumes are spatially and temporally correlated, which means that if the TM is low-rank or sparse in a transform domain, the TM is estimated using CS. In [24], based on the low-rank property of the TM and the fact that rows and columns close to each other have comparable values, the TM is estimated using sparsity-regularized matrix factorization. In [25], the TM is estimated based on the fact that the TM is sparse in a transform domain. On the other hand, in this study, we do not consider that the sequences of traffic volumes are sparse. Instead, we assume sparsity in the *connectivity* of VMs, meaning that only a few pairs of VMs transfer packets between them. Furthermore, to solve the optimization problem for estimating the flow traffic volumes, we consider additional constraints of measured traffic volumes at NICs and information of virtual LAN (VLAN) in the network.

3. VM Placement Method Using Compressed Sensing-Based Traffic Estimator

3.1 System Model

Figure 1 shows the network structure in this study, including N VMs, M PHs, and one switch. In this structure, the PHs are connected to the switch via their NICs. For simplicity, we assume that each PH has only one NIC. We define $\mathcal{V}^{(\mathrm{VM})} = \{v_n^{(\mathrm{VM})} \mid n = 1,2,\ldots,N\}$ as the set of VMs and $\mathcal{V}^{(\mathrm{NIC})} = \{v_m^{(\mathrm{NIC})} \mid m = 1,2,\ldots,M\}$ as the set of NICs. The set $\mathcal{V}^{(\mathrm{VM})}$ is divided into N_G *groups* $\mathcal{V}_i^{(\mathrm{VM})}$ $(i = 1,2,\ldots,N_\mathrm{G})$, where $\bigcup_i^{N_\mathrm{G}} \mathcal{V}_i^{(\mathrm{VM})} = \mathcal{V}^{(\mathrm{VM})}$ and $\mathcal{V}_i^{(\mathrm{VM})} \cap \mathcal{V}_{i'}^{(\mathrm{VM})} = \emptyset$ $(i \neq i')$. In the figure, VMs with the same color constitute a group. VMs in a group constitute a VLAN and we assume that each VM transmits and receives packets to and from VMs in the same group. We also define $\mathcal{H}_m \subset \mathcal{V}^{(\mathrm{VM})}$ $(m = 1,2,\ldots,M)$ as the set of VMs placed in the m-th PH, where $\bigcup_{m=1}^{M} \mathcal{H}_m = \mathcal{V}^{(\mathrm{VM})}$.

We refer to VMs and NICs as *nodes* and define $\mathcal{V} = \mathcal{V}^{(\mathrm{VM})} \cup \mathcal{V}^{(\mathrm{NIC})}$ as the set of all nodes. Each node can measure the traffic volumes transmitted from and received at the node. The pair (k,l) $(k,l = 1,2,\ldots,N, k \neq l)$ of VM $v_k^{(\mathrm{VM})} \in \mathcal{V}^{(\mathrm{VM})}$ and $v_l^{(\mathrm{VM})} \in \mathcal{V}^{(\mathrm{VM})}$ is referred to as the (k,l)-th *flow*. Since the number of flows is $N(N-1)$, it is not practical to measure the traffic volume of all flows, especially in a network with a large number of VMs. Therefore, we assume that the nodes cannot measure the traffic volume of each flow.

We consider discrete time t $(t = 1,2,\ldots)$ by dividing the continuous time with interval Δ $(\Delta > 0)$. $x_n^{(\mathrm{tx})}(t)$ and $x_{n'}^{(\mathrm{rx})}(t)$ $(n,n' = 1,2,\ldots,N)$ denote the traffic volumes transmitted from $v_n^{(\mathrm{VM})}$ and received at $v_{n'}^{(\mathrm{VM})}$ within the t-th time interval, respectively. We define $\boldsymbol{x}^{(\mathrm{tx})}(t) = (x_1^{(\mathrm{tx})}(t)\ x_2^{(\mathrm{tx})}(t)\ \cdots\ x_N^{(\mathrm{tx})}(t))^\top$ and $\boldsymbol{x}^{(\mathrm{rx})}(t) = (x_1^{(\mathrm{rx})}(t)\ x_2^{(\mathrm{rx})}(t)\ \cdots\ x_N^{(\mathrm{rx})}(t))^\top$. $y_m^{(\mathrm{out})}$ denotes the traffic volume sent from the m-th PH through $v_m^{(\mathrm{NIC})} \in \mathcal{V}^{(\mathrm{NIC})}$ within the t-th time interval, and $y_m^{(\mathrm{in})}$ denotes the traffic volume injected into the m-th PH through $v_m^{(\mathrm{NIC})}$. We define $\boldsymbol{y}^{(\mathrm{in})}(t) = (y_1^{(\mathrm{in})}(t)\ y_2^{(\mathrm{in})}(t)\ \cdots\ y_M^{(\mathrm{in})}(t))^\top$ and $\boldsymbol{y}^{(\mathrm{out})}(t) = (y_1^{(\mathrm{out})}(t)\ y_2^{(\mathrm{out})}(t)\ \cdots\ y_M^{(\mathrm{out})}(t))^\top$. We define *node traffic vector* $\boldsymbol{z}(t)$ at time t as

$$\boldsymbol{z}(t) = \begin{pmatrix} \boldsymbol{x}^{(\mathrm{tx})}(t) \\ \boldsymbol{x}^{(\mathrm{rx})}(t) \\ \boldsymbol{y}^{(\mathrm{in})}(t) \\ \boldsymbol{y}^{(\mathrm{out})}(t) \end{pmatrix}.$$

Let $u_{k,l}(t)$ $(k,l = 1,2,\ldots,N, k \neq l)$ denote the traffic volume transmitted from $v_k^{(\mathrm{VM})}$ and received at $v_l^{(\mathrm{VM})}$ within the t-th time interval. We define the *flow traffic vector* $\boldsymbol{u}(t) = (u_{1,2}(t)\ u_{1,3}(t)\ \cdots\ u_{N,N-1}(t))^\top$ at time t. We

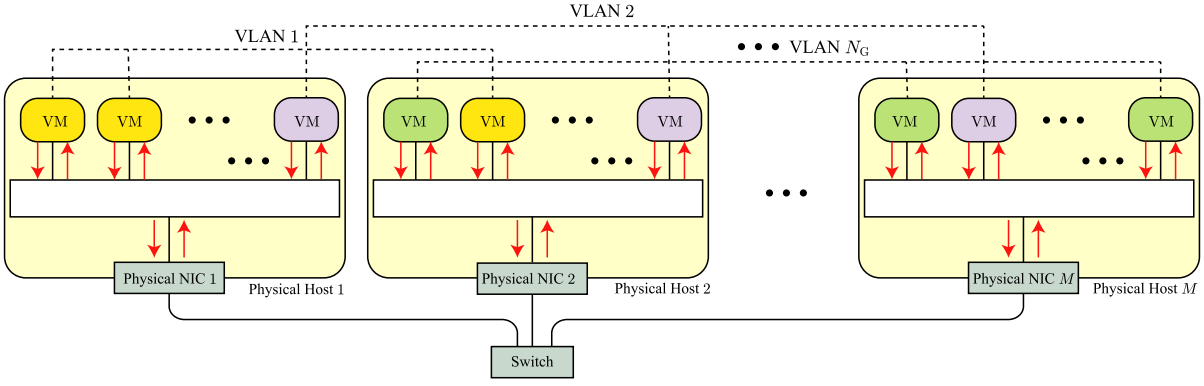


Fig. 1 Network configuration.

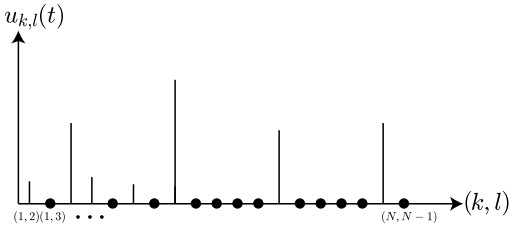


Fig. 2 Sparsity of flow traffic vector $\boldsymbol{u}$.

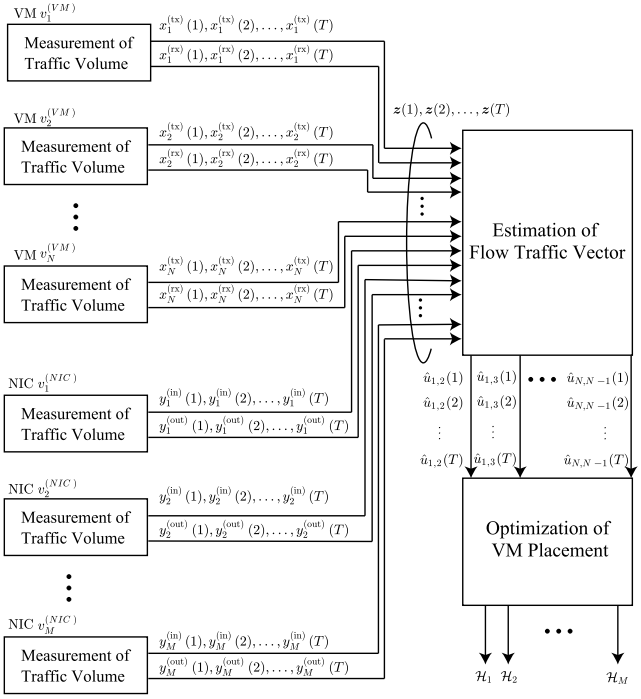


Fig. 3 Proposed VMP method.

assume that $\boldsymbol{u}(t)$ is approximately *sparse*, which means that only a few elements in $\boldsymbol{u}(t)$ have large values, as shown in Fig. 2. We consider this a natural assumption in practical network environments. In [27], traffic volumes of VM pairs measured in a datacenter network showed that only a few VM pairs have extreme amounts of traffic, validating the sparsity assumption. Furthermore, the sparsity increases as the number of groups increases. This is because we assume that VMs can send packets to other VMs in the same group.

3.2 Overview of the Proposed VMP Method

Figure 3 shows a block diagram of the proposed VMP method. We assume that $\boldsymbol{z}(t)$ $(t = 1,2,\ldots)$ is measurable at the VMs. The purpose of the proposed VMP method is to determine $\mathcal{H}_m$ $(m = 1,2,\ldots,M)$ from set $\mathcal{Z} = \{\boldsymbol{z}(t) \mid t = 1,2,\ldots,T\}$ of measured node traffic vectors, where T $(T > 0)$ denotes the *measurement period*. From $\mathcal{Z}$, a set $\hat{\mathcal{U}} = \{\hat{\boldsymbol{u}}(t) \mid t = 1,2,\ldots,T\}$ of estimated flow traffic vectors is obtained with CS, as explained in Sect. 3.3.

We refer to traffic within a PH as *PH traffic* of the host and traffic transmitted from inside to outside of the physical host as *NIC traffic* of the host. The proposed VMP method aims to balance traffic loads in PH traffic volumes and NIC traffic volumes. In Sect. 3.4, the load-balancing problem is formulated as a mixed-integer nonlinear optimization problem. From $\hat{\mathcal{U}}$, $\mathcal{H} = \{\mathcal{H}_m \mid m = 1,2,\ldots,M\}$ is determined by solving the optimization problem.

In this study, we consider a VMP method with batch processing, where each VM is assigned a PH on the basis of T measurements. If the VM placement is immediately changed in response to flow traffic dynamics, the computational time for solving traffic estimation and optimization problems is crucial, even in the case of batch processing. However, in this study, we do not consider such a time-critical VMP problem. The proposed method can be extended to a *live migration* method [28]–[30], where the VMs are dynamically relocated according to traffic volumes, because flow traffic volumes are estimated at every measurement time. In this case, computational time is a crucial performance metric, which will be studied in the future.

3.3 Estimation of Flow Traffic Vector

The relationship between $x_n^{(\mathrm{tx})}(t)$ $(n = 1,2,\ldots,N,\ t = 1,2,\ldots)$ and $\boldsymbol{u}(t)$ can be expressed as

$$x_n^{(\mathrm{tx})} = \sum_{j=1}^{N}\sum_{\substack{k=1\\ j\neq k}}^{N} a_{n,j,k}u_{j,k}(t) = \boldsymbol{a}_n^{\top}\boldsymbol{u}(t), \tag{1}$$

$$\boldsymbol{a}_n = (a_{n,1,2}\ a_{n,1,3}\ \cdots\ a_{n,N,N-1})^{\top}.$$

If $v_n^{(\mathrm{VM})} \in \mathcal{V}_i^{(\mathrm{VM})}$, then $a_{n,j,k}$ $(n,j,k = 1,2,\ldots,N, i \neq j)$ is given by

$$a_{n,j,k} = \begin{cases} 1 & \text{if } j = n,\ v_k^{(\mathrm{VM})} \in \mathcal{V}_i^{(\mathrm{VM})} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the relationship between $x_n^{(\mathrm{rx})}(t)$ and $\boldsymbol{u}(t)$ can be expressed as

$$x_n^{(\mathrm{rx})} = \sum_{j=1}^{N}\sum_{\substack{k=1\\ j\neq k}}^{N} b_{n,k,j}u_{k,j}(t) = \boldsymbol{b}_n^{\top}\boldsymbol{u}(t), \tag{2}$$

$$\boldsymbol{b}_n = (b_{n,1,2}\ b_{n,1,3}\ \cdots\ b_{n,N,N-1})^{\top}.$$

If $v_n^{(\mathrm{VM})} \in \mathcal{V}_i^{(\mathrm{VM})}$, then $b_{n,k,j}$ $(n,j,k = 1,2,\ldots,N, i \neq j)$ is given by

$$b_{n,k,j} = \begin{cases} 1 & \text{if } j = n,\ v_k^{(\mathrm{VM})} \in \mathcal{V}_i^{(\mathrm{VM})} \\ 0 & \text{otherwise.} \end{cases}$$

The relationship between $y_m^{(\mathrm{in})}(t)$ $(m = 1,2,\ldots,M,\ t = 1,2,\ldots)$ and $\boldsymbol{u}(t)$ can be expressed as

$$y_m^{(\mathrm{in})}(t) = \sum_{n=1}^{N}\sum_{\substack{k=1\\ k\neq n}}^{N} c_{m,k,n}u_{k,n}(t) = \boldsymbol{c}_m^{\top}\boldsymbol{u}(t), \tag{3}$$

$$\boldsymbol{c}_m = (c_{m,1,2}\ c_{m,1,3}\ \cdots\ c_{m,N,N-1})^{\top},$$

where $c_{m,k,n}$ is given by

$$c_{m,k,n} = \begin{cases} 1 & \text{if } n \in \mathcal{H}_m, k \in \mathcal{V}^{(\mathrm{VM})} \setminus \mathcal{H}_m \\ 0 & \text{otherwise} \end{cases}.$$

Similarly, the relationship between $y_m^{(\mathrm{out})}(t)$ and $\boldsymbol{u}(t)$ can be expressed as

$$y_m^{(\mathrm{out})}(t) = \sum_{n=1}^{N} \sum_{\substack{k=1 \\ k \neq n}}^{N} d_{m,n,k} u_{n,k}(t) = \boldsymbol{d}_m^\top \boldsymbol{u}(t), \tag{4}$$
$$\boldsymbol{d}_m = (d_{m,1,2}\ d_{m,1,3}\ \cdots\ d_{m,N,N-1})^\top,$$

where $d_{m,n,k}$ is given by

$$d_{m,n,k} = \begin{cases} 1 & \text{if } n \in \mathcal{H}_m, k \in \mathcal{V}^{(\mathrm{VM})} \setminus \mathcal{H}_m \\ 0 & \text{otherwise} \end{cases}.$$

From (1)–(4), the relationship between $\boldsymbol{z}$ and $\boldsymbol{u}(t)$ can be expressed as

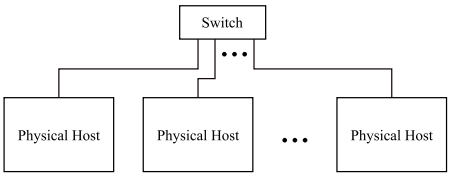
$$z(t) = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \\ \boldsymbol{C} \\ \boldsymbol{D} \end{pmatrix} \boldsymbol{u}(t) = \boldsymbol{R}\boldsymbol{u}(t), \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \\ \boldsymbol{C} \\ \boldsymbol{D} \end{pmatrix},$$
$$\boldsymbol{A} = (\boldsymbol{a}_1\ \boldsymbol{a}_2\ \cdots\ \boldsymbol{a}_N)^\top, \boldsymbol{B} = (\boldsymbol{b}_1\ \boldsymbol{b}_2\ \cdots\ \boldsymbol{b}_N)^\top,$$
$$\boldsymbol{C} = (\boldsymbol{c}_1\ \boldsymbol{c}_2\ \cdots\ \boldsymbol{c}_M)^\top, \boldsymbol{D} = (\boldsymbol{d}_1\ \boldsymbol{d}_2\ \cdots\ \boldsymbol{d}_M)^\top.$$

The problem of estimating flow traffic volumes can be viewed as a linear inverse problem to estimate $\boldsymbol{u}(t)$ from $\boldsymbol{z}(t)$. The lengths of $\boldsymbol{z}(t)$ and $\boldsymbol{u}(t)$ are $2(N + M)$ and $N(N - 1)$, respectively. Therefore, if $N > M$ and $N > 5$, the length of $\boldsymbol{u}(t)$ is greater than that of $\boldsymbol{z}(t)$ because $N(N - 1) - 2(N + M) > N^2 - 5N = N(N - 5) > 0$. This means that the problem is an underdetermined linear inverse problem, which means that there are an infinite number of solutions [21]. In this study, under the assumption that $\boldsymbol{u}(t)$ is a sparse vector, as described in Sect. 3.1, $\boldsymbol{u}(t)$ is estimated with CS from the measured node traffic vector $\boldsymbol{z}(t)$. In the proposed method, we obtain estimated flow traffic vector $\hat{\boldsymbol{u}}(t) = (\hat{u}_{1,2}(t)\ \hat{u}_{1,3}(t)\ \cdots\ \hat{u}_{N,N_1}(t))^\top$ of $\boldsymbol{u}(t)$ by solving the following ℓ_1-ℓ_2 optimization problem [31]:

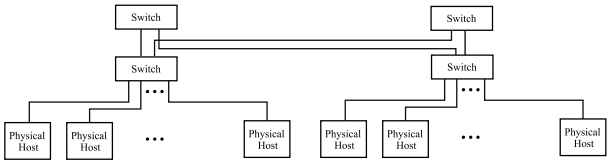
$$\min_{\boldsymbol{u}(t)} \quad \frac{1}{2}\|\boldsymbol{z}(t) - \boldsymbol{R}\boldsymbol{u}(t)\|_2^2 + \alpha\|\boldsymbol{u}(t)\|_1 \tag{5}$$
$$\text{subject to} \quad u_{i,j}(t) \geq 0, \quad i, j = 1, 2, \ldots, N, i \neq j.$$

In the proposed method, the optimization problem (5) is solved with CVX [32], a convex optimization library. α is the regularization parameter that weights the ℓ_1 regularization term $\|\boldsymbol{u}(t)\|_1$. As α increases, the sparsity of the estimated flow vector $\hat{\boldsymbol{u}}(t)$ becomes stronger, and only elements corresponding to flows with large traffic volumes in the estimated vector have nonzero values. Therefore, by setting α appropriately, only flows that have more impact on the network can be detected [33].

The formulation in the proposed flow traffic estimator is based on the relationship between VMs, that is, the connectivity between VMs and the physical hosts to which VMs are assigned. Therefore, although we consider the simple network topology in Fig. 4(a), the proposed method can be applied to general topologies, such as the Fat-Tree topology in Fig. 4(b), if the relationship is adequately formulated. The performance of the proposed method should be investigated in various environments. However, we do not evaluate the proposed method with such a complicated topology because our goal in this study is to evaluate its basic performance, and the evaluation in such an environment is beyond the scope of this study.



(a) Simple topology considered in this study.



(b) Example of Fat-Tree topology.

Fig. 4 Network topology.

3.4 VMP Optimization

To uniquely specify the hosts assigned to VMs, we define vectors $\boldsymbol{p}$ and $\boldsymbol{p}_m$ as

$$\boldsymbol{p} = (\boldsymbol{p}_1^\top\ \boldsymbol{p}_2^\top\ \cdots\ \boldsymbol{p}_M^\top)^\top$$
$$\boldsymbol{p}_m = (p_{m,1}\ p_{m,2}\ \cdots\ p_{m,N})^\top,$$

where $p_{m,n} = 1$ if $v_n^{(\mathrm{VM})} \in \mathcal{H}_m$ and $p_{m,n} = 0$ otherwise. We define $\boldsymbol{q} = (q_1\ q_2\ \boldsymbol{p}^\top)$ and $\boldsymbol{f} = (f_1\ f_2\ \boldsymbol{0}_{MN}^\top)^\top$. The optimum VMP is obtained by solving the following mixed-integer nonlinear programming problem $\mathcal{P}_1$:

$$\begin{aligned} \mathcal{P}_1 : \max_{\boldsymbol{q}} \quad & \boldsymbol{f}^\top \boldsymbol{q} \\ \text{subject to} \quad & q_1 \gamma_{\mathrm{all}} \boldsymbol{1}_M - \hat{\boldsymbol{U}}\boldsymbol{\phi}(\boldsymbol{p}) \leq \boldsymbol{0}_M \\ & q_2 \gamma_{\mathrm{all}} \boldsymbol{1}_M - \boldsymbol{U}\boldsymbol{\psi}(\boldsymbol{p}) \leq \boldsymbol{0}_M \\ & (\boldsymbol{0}_{N,2}\ \boldsymbol{E})\boldsymbol{q} = \boldsymbol{1}_{MN} \\ & 0 \leq q_1 \leq q_1^{(\max)}, \quad 0 \leq q_2 \leq q_2^{(\max)}, \\ & p_n^{(m)} \in \{0, 1\}, \\ & \quad n = 1, 2, \ldots, N, m = 1, 2, \ldots, M, \end{aligned}$$

where parameters $q_1^{(\max)}$ and $q_2^{(\max)}$ are set to $q_1^{(\max)} = 1$ and $q_2^{(\max)} = 1$, unless otherwise stated. The problem $\mathcal{P}_1$ is derived as shown in Appendix. In this study, we solve the optimization problem with the surrogate optimization solver

in MATLAB [34].

Although problem $\mathcal{P}_1$ can relocate all the VMs, it can be extended to relocate a subset of VMs. Let $\mathcal{H}_m^{(0)}$ $(m = 1,2,\ldots,M)$ denote the set of VMs assigned to PH m before applying the proposed VMP method. We define $\mathcal{V}_{\rm fix} = \{v_{i_k}^{(\rm VM)} \mid k = 1,2,\ldots,N_{\rm fix}, i_k \in \{1,2,\ldots,N\}\} \subset \mathcal{V}^{(\rm VM)}$ as a set of VMs that are not relocated by the proposed method and $\mathcal{I}_{\rm fix} = \{i_1,i_2,\ldots,i_{N_{\rm fix}} \mid v_{i_k}^{(\rm VM)} \in \mathcal{V}_{\rm fix}\}$ as the index set of VMs in $\mathcal{V}_{\rm fix}$. We also define $N_{\rm fix} \times N$ matrix $\boldsymbol{G}^{(m)} = [g_{k,l}^{(m)}]_{1\le k\le N_{\rm fix}, 1\le l\le N}$ as

$$g_{k,l}^{(m)} = \begin{cases} 1 & \text{if } i_k \in \mathcal{I}_{\rm fix}, v_{i_k}^{(\rm VM)} \in \mathcal{H}_m^{(0)}, l = i_k \\ 0 & \text{otherwise} \end{cases},$$

and $\boldsymbol{G} = (\boldsymbol{G}^{(1)}\ \boldsymbol{G}^{(2)}\ \cdots\ \boldsymbol{G}^{(M)}) \in \{0,1\}^{N_{\rm fix}\times MN}$. Then, $\boldsymbol{p}$ satisfies $\boldsymbol{G}\boldsymbol{p} = \boldsymbol{1}_{N_{\rm fix}}$, and problem $\mathcal{P}_1$ can be rewritten as

$$\begin{aligned} \mathcal{P}_2 : \max_{\boldsymbol{q}} \quad & \boldsymbol{f}^\top \boldsymbol{q} \\ \text{subject to} \quad & q_1\gamma_{\rm all}\boldsymbol{1}_M - \boldsymbol{U}\boldsymbol{\phi}(\boldsymbol{p}) \le \boldsymbol{0}_M \\ & q_2\gamma_{\rm all}\boldsymbol{1}_M - \boldsymbol{U}\boldsymbol{\psi}(\boldsymbol{p}) \le \boldsymbol{0}_M \\ & \begin{pmatrix} \boldsymbol{0}_{N,2} & \boldsymbol{E} \\ \boldsymbol{0}_{N_{\rm fix}} & \boldsymbol{G} \end{pmatrix} \boldsymbol{q} = \boldsymbol{1}_{MN} \\ & 0 \le q_1 \le q_1^{(\max)}, \quad 0 \le q_2 \le q_2^{(\max)}, \\ & p_n^{(m)} \in \{0,1\}, \\ & \quad n = 1,2,\ldots,N, m = 1,2,\ldots,M. \end{aligned}$$

$\mathcal{V}_{\rm fix}$ can be determined according to the estimated flow traffic vectors. In this study, we consider the following method based on flows with heavy traffic volumes. The (i,j)-th flow is referred to as a *heavy traffic flow* if $\sum_{t=1}^{T} \hat{u}_{i,j}(t)/T > u_{\rm th}$. $\overline{\mathcal{V}}_{\rm fix} \subseteq \mathcal{V}$ includes $v_i^{(\rm VM)}$ if $v_i^{(\rm VM)}$ has at least one heavy traffic flow, and is expressed as

$$\overline{\mathcal{V}}_{\rm fix} = \left\{ v_i^{(\rm VM)} \middle| \max_j \left\{ \sum_{t=1}^{T} \hat{u}_{i,j}(t)/T \right\} > u_{\rm th}, \right. \\ \left. i,j \in \{1,2,\ldots,N\}, i \ne j \right\}.$$

$\mathcal{V}_{\rm fix}$ is then given by $\mathcal{V}_{\rm fix} = \mathcal{V} \setminus \overline{\mathcal{V}}_{\rm fix}$. This method aims at relocating VMs that transmit heavy traffic volumes and is compatible with the CS-based flow traffic estimator, which estimates high flow traffic volumes by regarding other flow traffic volumes as zeros.

4. Performance Evaluation

4.1 Network Configuration

We evaluate the proposed VMP method with simulation experiments. Unless otherwise noted, we set the number N of VMs to $N = 30$ and the number M of PHs to $M = 3$, and the number $N_{\rm G}$ of groups to $N_{\rm G} = 3$. Each VM is assigned to one group randomly, and is assigned to one host so that all the hosts contain the same number of VMs. For $M = 3$, H_m $(m = 1,2,3)$ is given by

$$\begin{aligned} \mathcal{H}_1 =& \{v_1^{(\rm VM)}, v_4^{(\rm VM)}, v_7^{(\rm VM)}, v_{10}^{(\rm VM)}, v_{13}^{(\rm VM)}, v_{16}^{(\rm VM)}, v_{19}^{(\rm VM)}, \\ & v_{22}^{(\rm VM)}, v_{25}^{(\rm VM)}, v_{28}^{(\rm VM)}\} \\ \mathcal{H}_2 =& \{v_2^{(\rm VM)}, v_5^{(\rm VM)}, v_8^{(\rm VM)}, v_{11}^{(\rm VM)}, v_{14}^{(\rm VM)}, v_{17}^{(\rm VM)}, v_{20}^{(\rm VM)}, \\ & v_{23}^{(\rm VM)}, v_{26}^{(\rm VM)}, v_{29}^{(\rm VM)}\} \\ \mathcal{H}_3 =& \{v_3^{(\rm VM)}, v_6^{(\rm VM)}, v_9^{(\rm VM)}, v_{12}^{(\rm VM)}, v_{15}^{(\rm VM)}, v_{18}^{(\rm VM)}, v_{21}^{(\rm VM)}, \\ & v_{24}^{(\rm VM)}, v_{27}^{(\rm VM)}, v_{30}^{(\rm VM)}\}. \end{aligned}$$

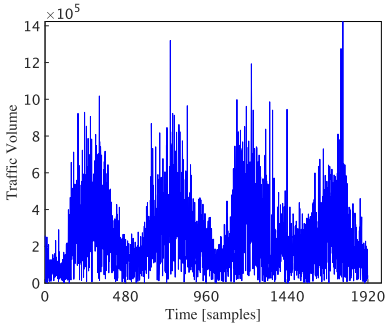
We set interval Δ between successive measurements to $\Delta = 3$ [min] and the measurement period to $T = 480$ [samples], which corresponds to one day.

4.2 Traffic Model

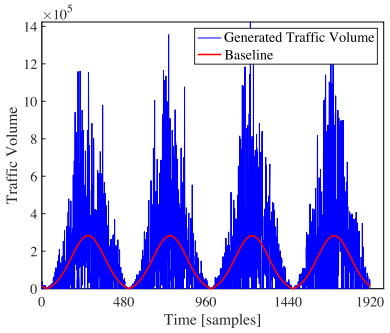
Sequences of traffic volumes are generated using a traffic model based on real traffic. Figure 5(a) shows an example of the volume of real flow traffic observed over a four-day period, which is measured in Fujitsu Ltd. The figure shows that the sequence of traffic volume fluctuates periodically, where each period has the highest value around the center of the period. From this observation, traffic volume is generated by a Gaussian-type baseline with Gaussian fluctuations.

We define the baseline $\mu(t)$ $(t = 1,2,\ldots,T)$ as

$$\mu(t) = F_{\rm peak}\left(\exp\left(-\frac{1}{2}\left(\frac{2\delta}{T}\right)^2\left(t-\frac{T}{2}\right)^2\right) - \mu_{\rm bias}\right), \\ t = 1,2,\ldots,T$$

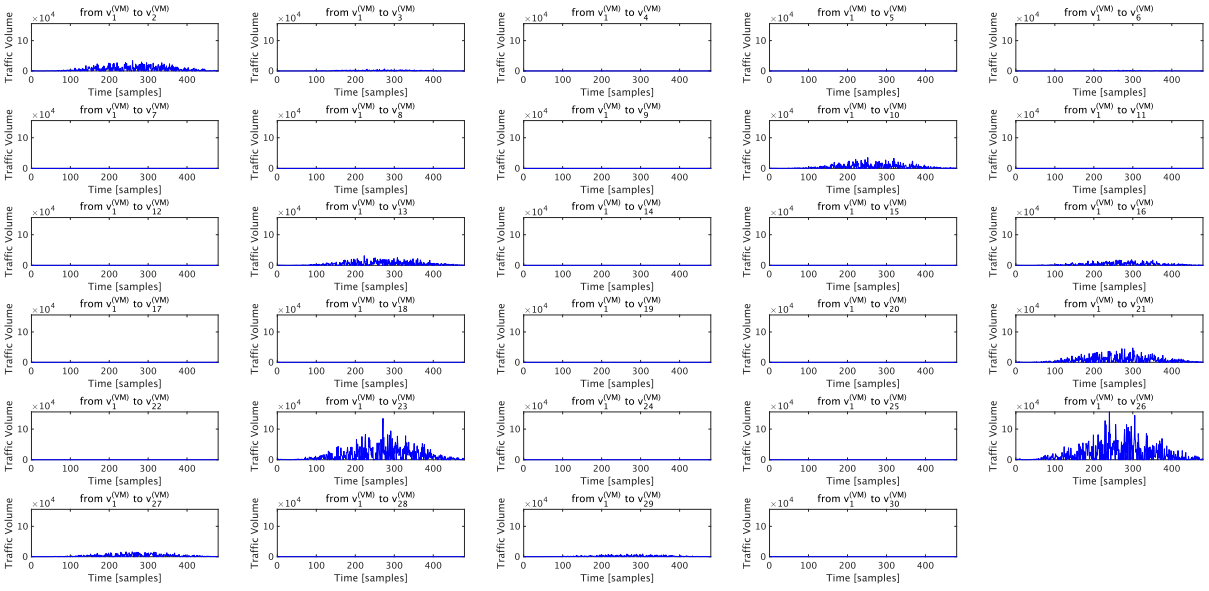


(a)

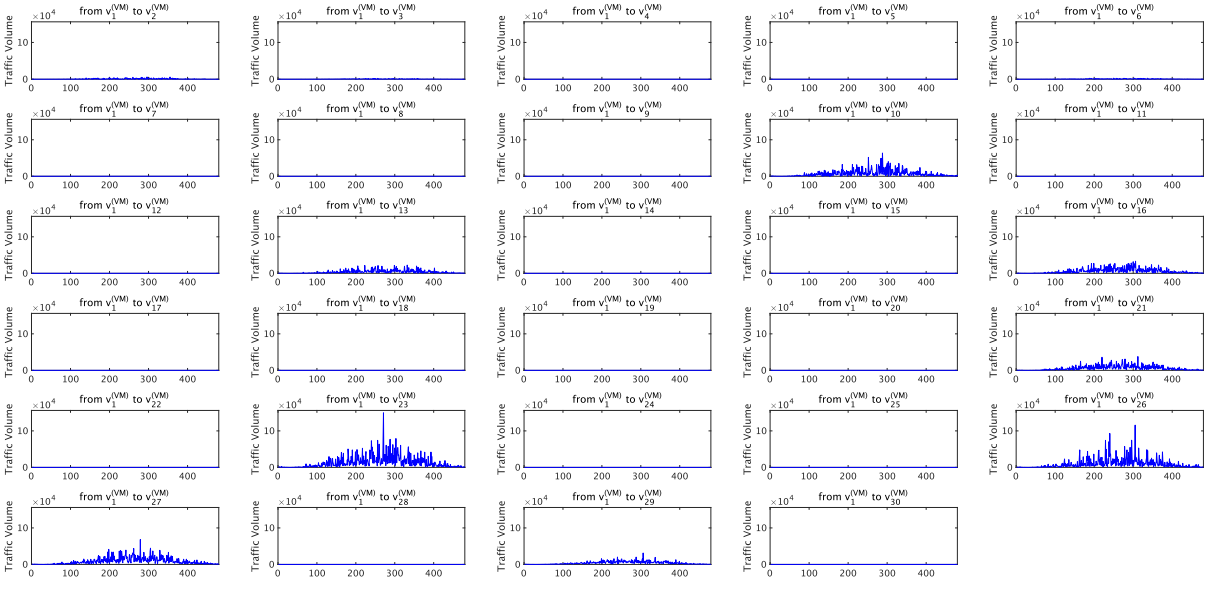


(b)

Fig. 5 Traffic model: (a) real traffic example, (b) generated traffic example.



(a)



(b)

Fig. 6 Traces of flow traffic volume transmitted from $v_1^{(\mathrm{VM})}$ ($\alpha = 1.0$): (a) true flow traffic traces, (b) estimated flow traffic traces.

where F_{peak} denotes the baseline peak value and δ denotes a parameter to determine the shape of the baseline. μ_{bias} denotes the bias parameter to determine the minimum value of the baseline. In this study, we set μ_{bias} to satisfy $\mu(0) = \mu(T) = 0$. The relationship among μ_{bias}, F_{peak}, and δ is then given by

$$\mu_{\mathrm{bias}} = F_{\mathrm{peak}} \exp(-\delta^2/2). \tag{6}$$

Let $u_{\mathrm{real}}(t)$ denote a sequence of real traffic volume. F_{peak} and δ are obtained by optimizing the following least-square problem:

$$\min_{F_{\mathrm{peak}},\delta} \quad \sum_{t=1}^{T} (u_{\mathrm{real}}(t) - u_{\mathrm{real}}(t))^2 \tag{7}$$

$$\text{subject to} \quad F_{\mathrm{peak}} \geq 0, \quad \delta \geq 0$$

By using $\mu(t)$, a flow traffic volume $u(t)$ ($t = 1, 2, \ldots, T$) is generated by

$$u(t) = \max\{\tilde{u}(t), 0\} \tag{8}$$
$$\tilde{u}(t) \sim \mathcal{N}\left(\mu_{\mathrm{base}}(t), (c\mu_{\mathrm{base}}(t))^2\right),$$

where c denotes the coefficient of variation; we set $c = 1.5$ in this study. Figure 5 shows an example of generated traffic volume. In the simulation experiments, δ is set to $\delta = 2.4637$, which is obtained by (7). Let $\mathcal{U}_{\mathrm{gen}} = \{u_{i,j}(t) \mid i, j = 1, 2, \ldots, N, i \neq j, t = 1, 2, \ldots, T\}$ denote a set of flow traffic volumes generated by (8). F_{peak} of each flow in $\mathcal{U}_{\mathrm{gen}}$ is

assigned by an exponential distribution with a mean of 10^4. To evaluate the proposed method, we generate $N_{\text{set}} = 100$ sets of flow traffic volumes.

4.3 Performance Evaluation of Traffic Estimation Method

We first evaluate the performance of the CS-based traffic estimator. Figure 6 shows a part of the traces of flow traffic volumes, where Figs. 6(a) and 6(b) correspond to the true traces and estimated traces transmitted from $v_1^{(\text{VM})}$, respectively. We have confirmed that similar results are obtained for other traces. From the figure, we can identify the flows with large traffic volume with the proposed method.

The proposed traffic estimator aims to identify flows with larger traffic volume rather than estimating the traces correctly. Therefore, to set parameters in the estimator, we evaluate it with *false positive rate* P_{FP} and *false negative rate* P_{FN}. We calculated P_{FP} and P_{FN} from true flow traffic volumes $u_{i,j}(t)$ and estimated flow traffic volumes $\hat{u}_{i,j}(t)$. We define average flow traffic volumes $\hat{\boldsymbol{u}} = (\hat{u}_{1,2}\ \hat{u}_{1,3}\ \ldots\ \hat{u}_{N,N-1})^\top$ for estimated flow traffic volumes and $\bar{\boldsymbol{u}} = (\bar{u}_{1,2}\ \bar{u}_{1,3}\ \ldots\ \bar{u}_{N,N-1})^\top$ for true flow traffic volumes as

$$\hat{u}_{i,j} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{i,j}(t), \quad \bar{u}_{i,j} = \frac{1}{T}\sum_{t=1}^{T}u_{i,j}(t).$$

For the k-th set of flow traffic volumes ($k = 1, 2, \ldots, N_{\text{set}}$), we define $\mathcal{I}_{\text{P}}$, $\mathcal{I}_{\text{N}}$, $\mathcal{I}_{\text{FP}}$, and $\mathcal{I}_{\text{FN}}$ as

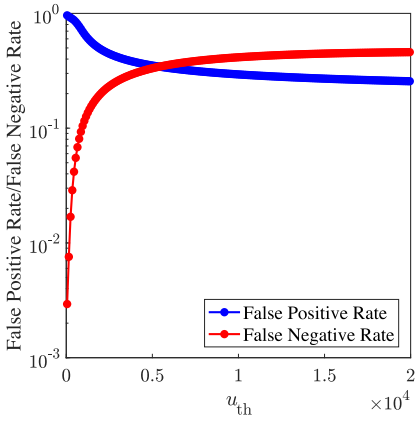
$$\begin{aligned}
\mathcal{I}_{\text{P}}^{(k)} &= \{(i,j) \mid \bar{u}_{i,j} \geq u_{\text{th}}, i, j = 1, 2, \ldots, N, i \neq j\} \\
\mathcal{I}_{\text{N}}^{(k)} &= \{(i,j) \mid \bar{u}_{i,j} < u_{\text{th}}, i, j = 1, 2, \ldots, N, i \neq j\} \\
\mathcal{I}_{\text{FP}}^{(k)} &= \{(i,j) \mid \hat{u}_{i,j} \geq u_{\text{th}}, \bar{u}_{i,j} < u_{\text{th}}, \\
&\qquad i, j = 1, 2, \ldots, N, i \neq j\}, \\
\mathcal{I}_{\text{FN}}^{(k)} &= \{(i,j) \mid \hat{u}_{i,j} < u_{\text{th}}, \bar{u}_{i,j} \geq u_{\text{th}}, \\
&\qquad i, j = 1, 2, \ldots, N, i \neq j\},
\end{aligned}$$

where u_{th} is the threshold parameter to idenfity heavy traffic flows, as described in Sect. 3.4. P_{FP} and P_{FN} are then defined as

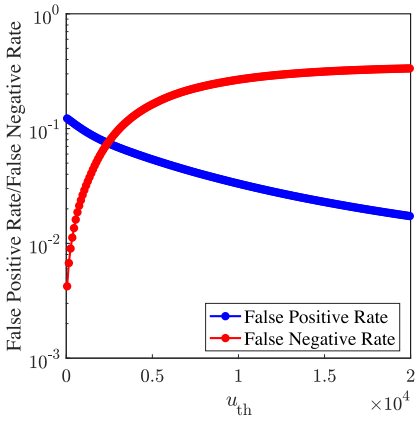
$$P_{\text{FP}} = \frac{1}{N_{\text{set}}}\sum_{k=1}^{N_{\text{set}}}\frac{|\mathcal{I}_{\text{FP}}^{(k)}|}{|\mathcal{I}_{\text{N}}^{(k)}|}, \quad P_{\text{FN}} = \frac{1}{N_{\text{set}}}\sum_{k=1}^{N_{\text{set}}}\frac{|\mathcal{I}_{\text{FN}}^{(k)}|}{|\mathcal{I}_{\text{P}}^{(k)}|}.$$

Figure 7 shows P_{FP} and P_{FN} vs. u_{th} for $\alpha = 0, 10^3, 10^5$. As shown in Figs. 7(a) and 7(b), P_{FP} and P_{FN} are higher when $\alpha = 0$. Because $\alpha = 0$ corresponds to the positively constrained least- squares problem, as can be seen from (5), this indicates that CS is effective in identifying heavy traffic flows. As shown in Fig. 7(c), P_{FNR} shows a higher value for excessively large α. The reason is that α represents the sparsity measure in the ℓ_1-ℓ_2 optimization problem [31], [33], and excessively large α suppresses most elements in the estimated flow traffic vector.

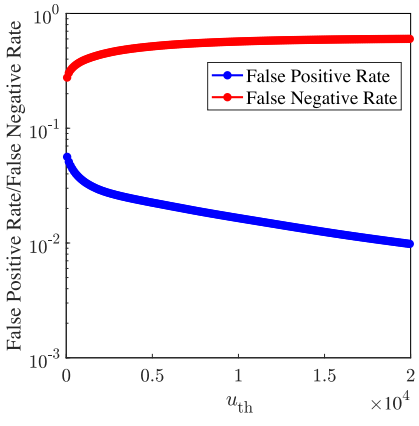
Figure 8 shows P_{FP} and P_{FN} vs. parameter α. We observe that P_{FP} and P_{FN} exhibit comparable values for $\alpha \in [10^{-3}, 10^2]$. In what follows, we focus on $\alpha = 1.0$. Figure 9 shows P_{FP} and P_{FN} for four different network configurations $(N, M, N_{\text{G}}) = (30,3,3), (30,5,3), (30,3,5), (50,3,3)$. N and N_{G} are important parameters that affect the performance of the proposed flow traffic estimator. We observe that both P_{FP} and P_{FN} in all the configurations exhibit smaller values for $\alpha = 1.0$. P_{FP} and P_{FN} in $(N, M, N_{\text{G}}) = (30,3,3), (30,5,3)$



(a)



(b)



(c)

Fig. 7 False positive rate R_{FP} and false negative rate R_{FN} vs. threshold u_{th} for α = (a) 0, (b) 10^3, and (c) 10^5.

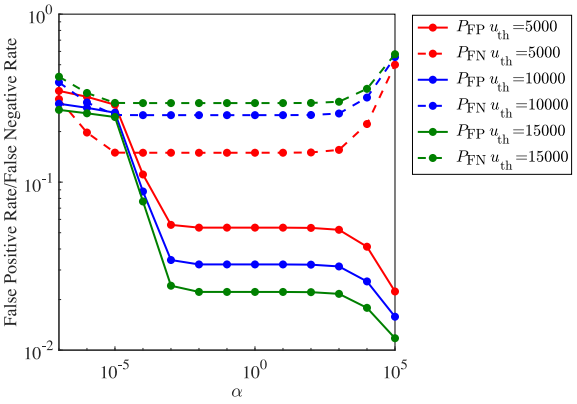


Fig. 8 False positive rate P_{FP} and false negative rate P_{FN} vs. parameter α.

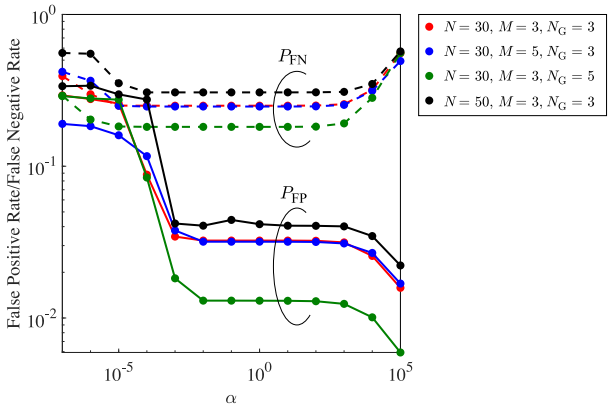


Fig. 9 False positive rate P_{FP} and false negative rate P_{FN} vs. parameter α for $u_{th} = 10000$ and $(N, M, N_G) = (30, 3, 3), (30, 5, 3), (30, 3, 5), (50, 3, 3)$.

are almost the same value when $\alpha = 1.0$ because the number M of hosts does not affect the sparsity of the flow traffic vector. P_{FP} and P_{FN} in $(N, M, N_G) = (30, 3, 5)$ are smaller than those in $(N, M, N_G) = (30, 3, 3)$. This is because the sparsity increases with the number N_G under the assumption that each VM transmits and receives packets to and from VMs in the same group, as described in Sect. 3.1. P_{FP} and P_{FN} in $(N, M, N_G) = (50, 3, 3)$ are larger than those in $(N, M, N_G) = (30, 3, 3)$. The reason is as follows. The number of elements in the flow traffic vector increases with $O(N^2)$. Therefore, larger N decreases the ability to estimate flow traffic volumes with compressed sensing. These results indicate that the proposed method may have worse performance in networks with an excessively large number of VMs. We will consider performance improvement in such an environment in future research. In the next subsection, we evaluate the effect of load balancing in the proposed method for $N = 30$, $M = 3$, and $N_G = 3$.

4.4 Performance Evaluation of the Proposed VMP Method

To evaluate the effect of load balancing in the proposed VMP method, we focus on a special case of the sets $\mathcal{V}_i$ $(i = 1, 2, \ldots, 5)$ of groups:

Table 2 Number $|\bar{\mathcal{V}}_{fix}|$ of heavy traffic nodes in Fig. 11.

u_{th}	5000	10000	15000
$\lvert\bar{\mathcal{V}}_{fix}\rvert$	20	13	7
$1 - \lvert\bar{\mathcal{V}}_{fix}\rvert/\lvert\mathcal{V}\rvert$	0.33	0.57	0.77

$$\mathcal{V}_1 = \{v_1^{(VM)}, v_2^{(VM)}, v_3^{(VM)}, v_6^{(VM)}, v_{10}^{(VM)}, v_{13}^{(VM)}, v_{16}^{(VM)}, v_{21}^{(VM)}, v_{23}^{(VM)}, v_{26}^{(VM)}, v_{27}^{(VM)}, v_{29}^{(VM)}\}$$
$$\mathcal{V}_2 = \{v_4^{(VM)}, v_7^{(VM)}, v_8^{(VM)}, v_9^{(VM)}, v_{14}^{(VM)}, v_{24}^{(VM)}, v_{28}^{(VM)}, v_{30}^{(VM)}\}$$
$$\mathcal{V}_3 = \{v_5^{(VM)}, v_{11}^{(VM)}, v_{12}^{(VM)}, v_{15}^{(VM)}, v_{17}^{(VM)}, v_{18}^{(VM)}, v_{19}^{(VM)}, v_{20}^{(VM)}, v_{22}^{(VM)}, v_{25}^{(VM)}\}$$

We evaluate the performance of the proposed VMP method for a set $\mathcal{U}_{gen}$ of flow traffic volumes by setting $f_1 = f_2 = 1.0$. Figure 10 shows PH traffic volume $\gamma_m^{(PH)}$ and NIC traffic volume $\gamma_m^{(PH)}$ $(m = 1, 2, 3)$ defined in (A· 2) and (A· 3), where $\gamma_m^{(PH)}$ represents the average traffic volume in PH m and $\gamma_m^{(NIC)}$ represents the average traffic volume sent through the NIC in the PH m to the outside of the PH. In Fig. 10(a), the proposed VMP method is not applied. Whereas $\mathcal{H}_m$ $(m = 1, 2, 3)$ are optimized by solving the problem $\mathcal{P}_1$ with true flow traffic vectors $\boldsymbol{u}(t)$ $(t = 1, 2, \ldots, T)$ as shown in Fig. 10(b), they are optimized with the estimated flow traffic vectors $\hat{\boldsymbol{u}}(t)$ $(t = 1, 2, \ldots, T)$ in Fig. 10(c). From the figures, we observe that the proposed VM placement method balances both PH traffic volumes and NIC traffic volumes.

Figure 11 shows $\gamma_m^{(PH)}$ and $\gamma_m^{(NIC)}$ when $\mathcal{H}_m$ $(m = 1, 2, 3)$ are optimized by solving the problem $\mathcal{P}_2$. For Figs. 11(a), 11(b), 11(c), u_{th} is set to 5000, 10000, and 15000, respectively. As shown in Figs. 10(c) and 11(a), the proposed method with $u_{th} = 5000$ exhibits a performance comparable to that with $u_{th} = 0$, which is obtained by the problem $\mathcal{P}_1$. However, when $u_{th} = 10000$ and 15000, the performance of the proposed method is degraded because the number of VMs available for relocation is reduced. Therefore, to maintain the load balancing capability, u_{th} should be set adequately. The numbers of heavy traffic nodes are summarized in Table 2. When $u_{th} = 5000$, the number of VMs relocated can be reduced by more than 30%.

Figures 12(a), 12(b) and 12(c) show the performance of the proposed method for $(q_1^{(max)}, q_2^{(max)}) = (1, 1), (0.05, 1), (1, 0.2)$, which is optimized with the problem $\mathcal{P}_1$. Comparing these figures, we observe that increasing $q_1^{(max)}$ results in increasing NIC traffic volumes and increasing $q_2^{(max)}$ results in increasing PH traffic volumes. Because $q_1^{(max)}$ represents the ratio of the minimum average PH traffic volume to the total traffic volume, the NIC traffic volumes can be increased by reducing $q_1^{(max)}$. However, because $q_2^{(max)}$ represents the ratio of the minimum average NIC traffic volume to the total traffic volume, the PH traffic volumes can be increased by reducing $q_2^{(max)}$. Figure 12 shows an example of a simulation result obtained using the simple method to adjust the balance

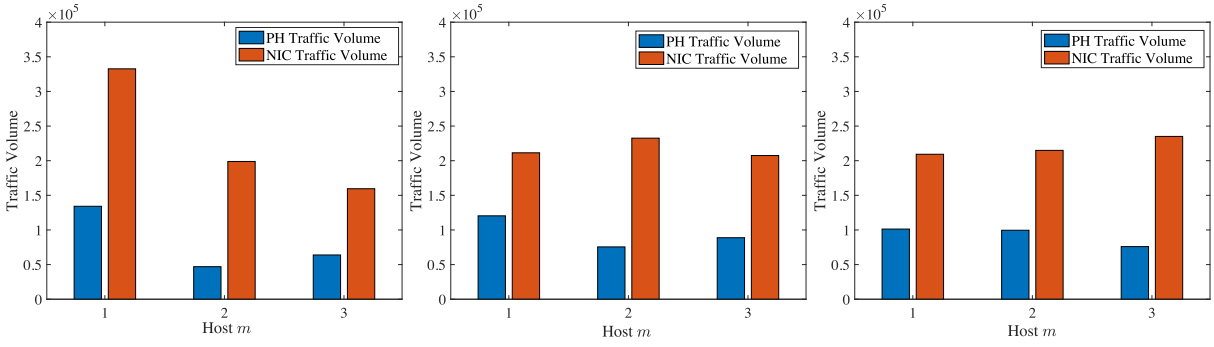


Fig. 10 Average traffic volumes with and without the proposed VM placement method: (a) Without the proposed VMP method, (b) with the proposed VMP method for true traffic flow volumes, and (c) with the proposed VMP method for estimated traffic flow volumes.

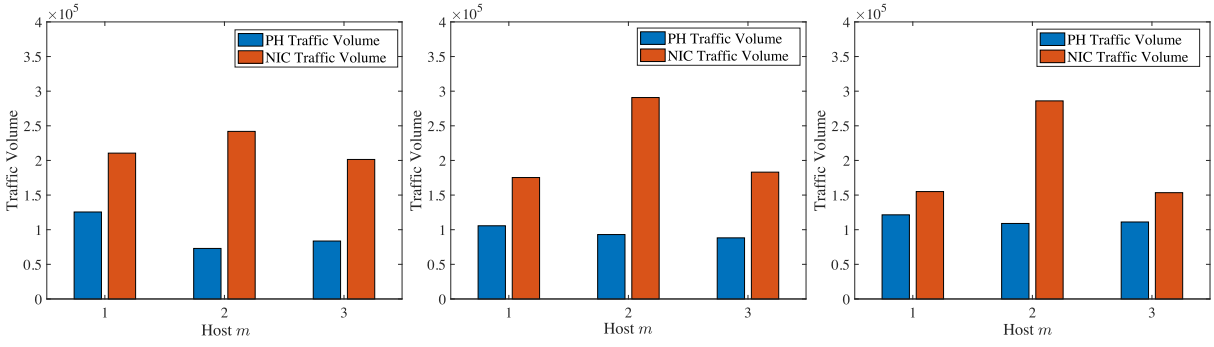


Fig. 11 Average traffic volumes with the proposed VM placement method for u_{th} = (a) 5000, (b) 10000, and (c) 15000.

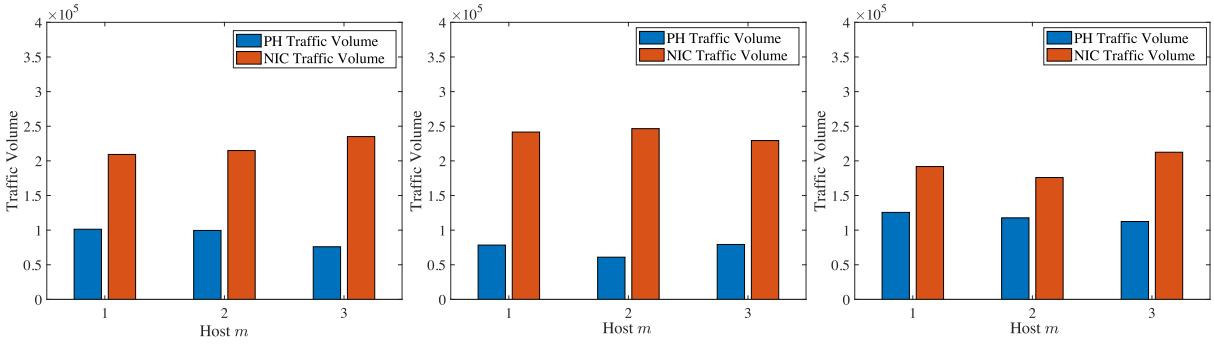


Fig. 12 Average traffic volumes with the proposed VM placement method for $(q_1^{(\max)}, q_2^{(\max)})$ = (a) (1, 1), (b) (0.05, 1) and (c) (1, 0.2). (a) is the same as Fig. 10(c).

between NIC and PH traffic volumes. Therefore, parameter optimization should be investigated in future work.

5. Conclusion

In this paper, we propose a VMP method based on the CS-based flow traffic estimator for load balancing of traffic volumes in PHs and NICs. In the flow traffic estimator, the relationship between node traffic volumes and flow traffic volumes is formulated with a system of linear equations, and the flow traffic volumes are estimated by solving the ℓ_1-ℓ_2 optimization problem. From the estimated flow traffic volumes, each VM is assigned to a PH by solving the mixed-integer optimization problem.

In the proposed VMP method, there are several remaining issues, which will be studied in future work.

- In the proposed method, PHs assigned to VMs are optimized based on *batch processing*, that is, the optimization problem is solved with flow traffic volumes estimated in T time instants. To apply the proposed method to live migration, the traffic estimator and the optimization problem should be extended to sequential processing.
- We have considered the VMP problem on the basis of traffic volume. However, in some network environments, it is also necessary to perform relocation taking into account various VM resources, such as CPU utilization and memory utilization, as well as traffic volume.

References

[1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vol.25, no.6, pp.599–616, June 2009.
[2] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," J. Internet Serv. Appl., vol.1, pp.7–18, April 2010.
[3] M. Masdari, S.S. Nabavi, and V. Ahmadi, "An overview of virtual machine placement schemes in cloud computing," Journal of Network and Computer Applications, vol.66, pp.106–127, May 2016.
[4] J. Xu and J.A.B. Fortes, "Multi-objective virtual machine placement in virtualized data center environments," Proc. 2010 IEEE/ACM International Conference on Green Computing and Communications (Green-Com) & International Conference on Cyber, Physical and Social Computing (CPSCom), pp.179–88, Dec. 2010.
[5] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," Computer Networks, vol.57, no.1, pp.179–196, Jan. 2013.
[6] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," Journal of Computer and System Sciences, vol.79, no.8, pp.1230–1242, Dec. 2013.
[7] X. Li, J. Wu, S. Tang, and S. Lu, "Let's stay together: Towards traffic aware virtual machine placement in data centers," Proc. IEEE INFOCOM 2014, pp.1842–1850, May 2014.
[8] M. Tang and S. Pan, "A hybrid genetic algorithm for the energy-efficient virtual machine placement problem in data centers," Neural Process Lett., vol.41, no.2, pp.211–221, April 2015.
[9] A.R. Ilkhechi, I. Korpeoglu, and ÖUlusoy, "Network-aware virtual machine placement in cloud data centers with multiple traffic-intensive components," Computer Networks, vol.91, pp.508–527, Nov. 2015.
[10] Q. Zheng, R. Li, X. Li, N. Shah, J. Zhang, F. Tian, K.-M. Chao, and J. Li, "Virtual machine consolidated placement based on multi-objective biogeography-based optimization," Future Generation Computer Systems, vol.54, pp.95–122, Jan. 2016.
[11] M. Mishra and U. Bellur, "Whither tightness of packing? The case for stable VM placement," IEEE Trans. Cloud Comput., vol.4, no.4, pp.481–494, Oct.-Dec. 2016.
[12] A. Choudhary, S. Rana, and K.J. Matahai, "A critical analysis of energy efficient virtual machine placement techniques and its optimization in a cloud computing environment," Procedia Computer Science, vol.78, pp.132–138, Dec. 2016.
[13] Y. Qin, H. Wang, F. Zhu, and L. Zhai, "A multi-objective ant colony system algorithm for virtual machine placement in traffic intense data centers," IEEE Access, vol.6, pp.58912–58923, Oct. 2018.
[14] A.K. Kulkarni and B. Annappa, "Context aware VM placement optimization technique for heterogeneous IaaS cloud," IEEE Access, vol.7, pp.89702–89713, July 2019.
[15] A. Ghasemi and A.T. Haghighat, "A multi-objective load balancing algorithm for virtual machine placement in cloud data centers based on machine learning," Computing, vol.102, no.9, pp.2049–2072, May 2020.
[16] H. Zhao, Q. Wang, J. Wang, B. Wan, and S. Li, "VM performance maximization and PM load balancing virtual machine placement in cloud," Proc. 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), May 2020.
[17] B. Zhang, X. Wang, and H. Wang, "Virtual machine placement strategy using cluster-based genetic algorithm," Neurocomputing, vol.428, pp.310–316, March 2021.
[18] S.M. Seyyedsalehi and M. Khansari, "Virtual machine placement optimization for big data applications in cloud computing," IEEE Access, vol.10, pp.96112–96127, Aug. 2022.
[19] R. Keshri and D.P. Vidyarthi, "Communication-aware, energy-efficient VM placement in cloud data center using ant colony optimization," Int. J. Inf. Tecnol., vol.15, no.8, pp.4529–4535, Oct. 2023.
[20] A. Yamamoto, K. Yumoto, T. Matsuda, J. Higuchi, T. Kodama, H. Ueno, and T. Shiraishi, "Virtual machine placement method with compressed sensing-based traffic volume estimation," Proc. IEEE International Conference on Consumer Electronics - Taiwan 2023 (ICCE-TW 2023), July 2023.
[21] K. Hayashi, M. Nagahara, and T. Tanaka, "A user's guide to compressed sensing for communications systems," IEICE Trans. Commun., vol.E96-B, no.3, pp.685–712, March 2013.
[22] E.J. Candès and M.B. Wakin, "An introduction to compressive sampling," IEEE Signal Process. Mag., vol.25, no.2, pp.21–30, March 2008.
[23] Y. Ohsita, T. Miyamura, S. Arakawa, S. Ata, E. Oki, K. Shiomoto, and M. Murata, "Gradually reconfiguring virtual network topologies based on estimated traffic matrices," IEEE/ACM Trans. Netw., vol.18, no.1, pp.177–189, Feb. 2010.
[24] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and Internet traffic matrices (extended version)," IEEE/ACM Trans. Netw., vol.20, no.3, pp.662–676, June 2012.
[25] D. Jiang, W. Wang, L. Shi, and H. Song, "A compressive sensing-based approach to end-to-end network traffic reconstruction," IEEE Trans. Netw. Sci. Eng., vol.7, no.1, pp.507–519, Jan. 2020
[26] Y. Tian, W. Chen, and C.-T. Lea, "An SDN-based traffic matrix estimation framework," IEEE Trans. Netw. Service Manag., vol.15, no.4, pp.1435–1445, Dec. 2018.
[27] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," Proc. IEEE INFOCOM 2010, March 2010.
[28] C. Clark, K. Fraser, S. Hand, J.G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," Proc. 2nd Symposium on Networked Systems Design and Implementation (NSDI 2005), pp.273–286, May 2005.
[29] H. Liu, C.-Z. Xu, H. Jin, J. Gong, and X. Liao, "Performance and energy modeling for live migration of virtual machines," Proc. 20th International Symposium on High Performance Distributed Computing, pp.171–182, June 2011.
[30] T.Z. He, A.N. Toosi, and R. Buyya, "Performance evaluation of live virtual machine migration in SDN-enabled cloud data centers," Journal of Parallel and Distributed Computing, vol.131, pp.55–68, Sept. 2019.
[31] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," IEEE Signal Process. Mag., vol.27, no.3, pp.76–88, May 2010.
[32] CVX Resarch, Inc., "CVX: Matlab software for disciplined convex programming," CVX Research, http://cvxr.com/cvx/
[33] T. Matsuda, M. Nagahara, and K. Hayashi, "Link quality classifier with compressed sensing based on ℓ_1-ℓ_2 optimization," IEEE Commun. Lett., vol.15, no.10, pp.1117–1119, Oct. 2011.

[34] MATLAB, https://www.mathworks.com/

Appendix: Derivation of Problem $\mathcal{P}_1$

Because each VM belongs to only one host, $p_{m,n}$ satisfies

$$\sum_{m=1}^{M} p_{m,n} = 1 \tag{A·1}$$

Let $\boldsymbol{I}_{N,N}$ denote the $N \times N$ identity matrix. Introducing $\boldsymbol{E} = (\boldsymbol{I}_{N,N}\ \boldsymbol{I}_{N,N}\ \cdots\ \boldsymbol{I}_{N,N}) \in \{0,1\}^{N\times MN}$, this can be rewritten as

$$\boldsymbol{E}\boldsymbol{p} = \boldsymbol{1}_N,$$

where $\boldsymbol{1}_N$ denotes the vector with length of N such that all elements are 1.

Let $\hat{\boldsymbol{\gamma}}^{(\mathrm{PH})} = (\hat{\gamma}_1^{(\mathrm{PH})}\ \hat{\gamma}_2^{(\mathrm{PH})}\ \cdots\ \hat{\gamma}_M^{(\mathrm{PH})})^\top$ and $\hat{\boldsymbol{\gamma}}^{(\mathrm{NIC})} = (\hat{\gamma}_1^{(\mathrm{NIC})}\ \hat{\gamma}_2^{(\mathrm{NIC})}\ \cdots\ \hat{\gamma}_M^{(\mathrm{NIC})})^\top$ denote the *PH traffic vector* and *NIC traffic vector*, respectively, where $\hat{\gamma}_m^{(\mathrm{PH})}$ and $\hat{\gamma}_m^{(\mathrm{NIC})}$ represent the average PH traffic volume and average NIC traffic volume of PH m, respectively, calculated from the set of estimated flow vectors $\{\hat{\boldsymbol{u}}(t) \mid t = 1,2,\ldots,T\}$. We define $\phi_{i,j}^{(m)}$ and $\psi_{i,j}^{(m)}$ for $m = 1,2,\ldots,M$ and $i,j = 1,2,\ldots,N$ as

$$\phi_{i,j}^{(m)} = \begin{cases} 1 & \text{if } v_i^{(\mathrm{VM})} \in \mathcal{H}_m, v_j^{(\mathrm{VM})} \in \mathcal{H}_m, i \neq j \\ 0 & \text{otherwise} \end{cases},$$

$$\psi_{i,j}^{(m)} = \begin{cases} 1 & \text{if } v_i^{(\mathrm{VM})} \in \mathcal{H}_m, v_j^{(\mathrm{VM})} \in \mathcal{V}^{(\mathrm{VM})} \setminus \mathcal{H}_m \\ 0 & \text{otherwise} \end{cases}.$$

$\hat{\gamma}_m^{(\mathrm{PH})}$ and $\hat{\gamma}_m^{(\mathrm{NIC})}$ are expressed as

$$\hat{\gamma}_m^{(\mathrm{PH})} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \phi_{i,j}^{(m)}\hat{u}_{i,j}(t), \tag{A·2}$$

$$\hat{\gamma}_m^{(\mathrm{NIC})} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \psi_{i,j}^{(m)}\hat{u}_{i,j}(t). \tag{A·3}$$

We define two $N \times N$ matrices $\boldsymbol{\Phi}^{(m)}(\boldsymbol{p}_m) = [\phi_{i,j}^{(m)}]_{1\le i,j\le N}$ and $\boldsymbol{\Psi}^{(m)}(\boldsymbol{p}_m) = [\psi_{i,j}^{(m)}]_{1\le i,j\le N}$ as functions of $\boldsymbol{p}^{(m)}$. $\boldsymbol{\Phi}^{(m)}$ and $\boldsymbol{\Psi}^{(m)}$ can be expressed as

$$\boldsymbol{\Phi}^{(m)} = \boldsymbol{p}_m\boldsymbol{p}_m^\top, \quad \boldsymbol{\Psi}^{(m)} = \boldsymbol{p}_m(\boldsymbol{1}_N - \boldsymbol{p}_m)^\top.$$

By vectorizing $\boldsymbol{\Phi}^{(m)}$ and $\boldsymbol{\Psi}^{(m)}$ without the diagonal elements, we obtain $\boldsymbol{\phi}^{(m)}$ and $\boldsymbol{\psi}^{(m)}$:

$$\boldsymbol{\phi}^{(m)}(\boldsymbol{p}_m) = (\phi_{1,2}^{(m)}\ \phi_{1,3}^{(m)}\ \cdots\ \phi_{N,N-1}^{(m)})^\top$$
$$\boldsymbol{\psi}^{(m)}(\boldsymbol{p}_m) = (\psi_{1,2}^{(m)}\ \psi_{1,3}^{(m)}\ \cdots\ \psi_{N,N-1}^{(m)})^\top$$

The PH traffic vector $\boldsymbol{\gamma}^{(\mathrm{PH})}$ and NIC traffic vector $\boldsymbol{\gamma}^{(\mathrm{NIC})}$ are expressed as

$$\hat{\boldsymbol{\gamma}}^{(\mathrm{PH})} = \hat{\boldsymbol{U}}\boldsymbol{\phi}(\boldsymbol{p}), \quad \hat{\boldsymbol{\gamma}}^{(\mathrm{NIC})} = \hat{\boldsymbol{U}}\boldsymbol{\psi}(\boldsymbol{p}),$$

$$\hat{\boldsymbol{U}} = \begin{pmatrix} \hat{\boldsymbol{u}}^\top & \boldsymbol{0}_{N(N-1)}^\top & \cdots & \boldsymbol{0}_{N(N-1)}^\top \\ \boldsymbol{0}_{N(N-1)}^\top & \hat{\boldsymbol{u}}^\top & \cdots & \boldsymbol{0}_{N(N-1)}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{N(N-1)}^\top & \boldsymbol{0}_{N(N-1)}^\top & \cdots & \hat{\boldsymbol{u}}^\top \end{pmatrix},$$

$$\hat{\boldsymbol{u}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{u}}(t),$$

$$\boldsymbol{\phi}(\boldsymbol{p}) = \begin{pmatrix} \boldsymbol{\phi}^{(1)}(\boldsymbol{p}_1) \\ \boldsymbol{\phi}^{(2)}(\boldsymbol{p}_2) \\ \vdots \\ \boldsymbol{\phi}^{(M)}(\boldsymbol{p}_M) \end{pmatrix}, \ \boldsymbol{\psi}(\boldsymbol{p}) = \begin{pmatrix} \boldsymbol{\psi}^{(1)}(\boldsymbol{p}_1) \\ \boldsymbol{\psi}^{(2)}(\boldsymbol{p}_2) \\ \vdots \\ \boldsymbol{\psi}^{(M)}(\boldsymbol{p}_M) \end{pmatrix}$$

Let γ_{all} denote the average traffic volume calcuated from all the flow traffic volume. From the definitions of $\gamma_m^{(\mathrm{PH})}$ and $\gamma_m^{(\mathrm{NIC})}$, we obtain the following relationship:

$$\begin{aligned}\gamma_{\mathrm{all}} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\hat{u}_{i,j}(t) \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{i\in\mathcal{H}_m}\sum_{\substack{j\in\mathcal{H}_m\\ j\neq i}}\hat{u}_{i,j}(t) \\ &\quad + \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{i\in\mathcal{H}_m}\sum_{\substack{j\in\mathcal{V}^{(\mathrm{VM})}\setminus\mathcal{H}_m\\ j\neq i}}\hat{u}_{i,j}(t) \\ &= \sum_{m=1}^{M}\gamma_m^{(\mathrm{PH})} + \sum_{m=1}^{M}\gamma_m^{(\mathrm{NIC})}\end{aligned}$$

Because $\gamma_m^{(\mathrm{PH})} \ge 0$ and $\gamma_m^{(\mathrm{NIC})} \ge 0$ for $\forall m \in \{1,2,\ldots,M\}$, $\gamma_m^{(\mathrm{PH})}$ and $\gamma_m^{(\mathrm{NIC})}$ always satisfy $\gamma_m^{(\mathrm{PH})} < \gamma_{\mathrm{all}}$ and $\gamma_m^{(\mathrm{NIC})} < \gamma_{\mathrm{all}}$, regardless of the VM placement $\{\mathcal{H}_m \mid m = 1,2,\ldots,M\}$.

Let $S(\boldsymbol{\gamma}^{\mathrm{PH}})$ and $S(\boldsymbol{\gamma}^{\mathrm{NIC}})$ denote variances of PH traffic volumes and NIC traffic volumes:

$$S(\boldsymbol{\gamma}^{(\mathrm{PH})}) = \frac{1}{M}\sum_{m=1}^{M}\left(\gamma_m^{(\mathrm{PH})} - \bar{\gamma}^{(\mathrm{PH})}\right)^2,$$

$$S(\boldsymbol{\gamma}^{(\mathrm{NIC})}) = \frac{1}{M}\sum_{m=1}^{M}\left(\gamma_m^{(\mathrm{NIC})} - \bar{\gamma}^{(\mathrm{NIC})}\right)^2,$$

$$\bar{\gamma}^{(\mathrm{PH})} = \frac{1}{M}\sum_{m=1}^{M}\gamma_m^{(\mathrm{PH})}, \ \bar{\gamma}^{(\mathrm{NIC})} = \frac{1}{M}\sum_{m=1}^{M}\gamma_m^{(\mathrm{NIC})}.$$

The proposed VMP method aims to reduce $S(\boldsymbol{\gamma}^{(\mathrm{PH})})$ and $S(\boldsymbol{\gamma}^{(\mathrm{NIC})})$ by appropriately assigning a PH to each VM. To do so, we adopt *max-min optimization*, where the variances are reduced by maximizing the minimum PH and NIC traffic volumes. By introducing slack variables q_1 and q_2 ($0 \le q_1, q_2 \le 1$), and using the property of $\gamma_m^{(\mathrm{PH})}$ and

$\gamma_m^{(\mathrm{NIC})}$ discussed above, the constraints for the max-min optimization problem are written as

$$\gamma_m^{(\mathrm{PH})} > q_1 \gamma_{\mathrm{all}}, \quad m = 1, 2, \ldots, M \tag{A·4}$$

$$\gamma_m^{(\mathrm{NIC})} > q_2 \gamma_{\mathrm{all}}, \quad m = 1, 2, \ldots, M \tag{A·5}$$

$S(\boldsymbol{\gamma}^{(\mathrm{PH})})$ and $S(\boldsymbol{\gamma}^{(\mathrm{NIC})})$ are minimized by maximizing q_1 and q_2. We define $\boldsymbol{q} = (q_1\ q_2\ \boldsymbol{p}^\top)^\top$ and $\boldsymbol{f} = (1\ 1\ \mathbf{0}_{MN}^\top)^\top$. From (A·1), (A·4), and (A·5), the optimization problem is formulated as the following mixed-integer nonlinear optimization problem:

$$\begin{aligned}
\max_{\boldsymbol{q}} \quad & \boldsymbol{f}^\top \boldsymbol{q} \\
\text{subject to} \quad & q_1 \gamma_{\mathrm{all}} \mathbf{1}_M - \hat{\boldsymbol{U}} \boldsymbol{\phi}(\boldsymbol{p}) \leq \mathbf{0}_M \\
& q_2 \gamma_{\mathrm{all}} \mathbf{1}_M - \hat{\boldsymbol{U}} \boldsymbol{\psi}(\boldsymbol{p}) \leq \mathbf{0}_M \\
& [\mathbf{0}_{N,2}\ \boldsymbol{E}] \boldsymbol{q} = \mathbf{1}_N \\
& 0 \leq q_1, q_2 \leq 1, \\
& p_n^{(m)} \in \{0, 1\}, \\
& \quad n = 1, 2, \ldots, N, m = 1, 2, \ldots, M.
\end{aligned}$$

Kenta Yumoto received his B.E. and M.S. degrees in engineering from Tokyo Metropolitan University in 2021, 2023, respectively. In 2023, he joined NTT Docomo Inc., Japan. His research includes traffic estimation techniques for cloud computing.

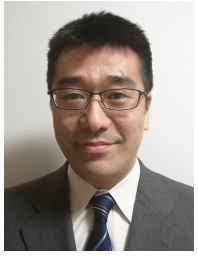
Ami Yamamoto received the B.E. degree from Tokyo Metropolitan University in 2023. She is currently a graduate student in the Graduate School of Systems Design, Tokyo Metropolitan University. Her research includes traffic estimation techniques for cloud computing.

Takahiro Matsuda received his B.E. with honors, M.E., and Ph.D. degrees in communications engineering from Osaka University in 1996, 1997, 1999, respectively. He joined the Department of Communications Engineering at the Graduate School of Engineering, Osaka University in 1999. In the same department, he was an Assistant Professor from 1999 to 2005, a Lecturer from 2005 to 2009, and an Associate Professor from 2009 to 2018. He is currently a Professor in the Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University. His research interests include performance analysis and the design of communication networks and wireless communications. He received the Best Tutorial Paper Award and the Best Magazine Paper Award from IEICE ComSoc in 2012, and the Best Paper Awards from IEICE in 2014 and 2023. He is a member of IPSJ and IEEE.

Junichi Higuchi received his B.E. and M.S. degrees in engineering from Tohoku University in 2006, 2008, respectively. In 2008, he joined Fujitsu Laboratories Ltd., Japan. He is currently a researcher with Fsas Technologies Inc. His research interests include IT operations analytics and AIOps for virtualization infrastructure.

Takeshi Kodama received his B.E. and M.S. degrees in science and engineering from Osaka University in 1998, 2000, respectively. In 2002, he joined Fujitsu Laboratories Ltd., Japan. His research interests include IT operation analytics and AIOps for virtualization infrastructure.

Hitoshi Ueno received his B.E. and M.S. degrees in communication engineering from Osaka University in 1995, 1997, respectively. In 1997, he joined Fujitsu Laboratories Ltd., Japan. He is currently a Manager with Fsas Technologies Inc. His research interests include IT operation analytics and AIOps for virtualization infrastructure.



Takashi Shiraishi received his B.E. and M.S. degrees in science and engineering from Tsukuba University in 2000, 2002, respectively. In 2002, he joined Fujitsu Laboratories Ltd., Japan. He is currently a Manager with Fujitsu Limited. His research interests include IT operation analytics and AIOps for virtualization infrastructure.

PAPER

Evaluation of Interference between 300 GHz Band Fronthaul Links Using Measured High Gain Antenna Radiation Patterns

Ken WATANABE[†], **Ryo OKUMURA**[†], ***Nonmembers***, **Akihiko HIRATA**[†a)], ***Senior Member***, ***and*** **Thomas KÜRNER**[††], ***Nonmember***

SUMMARY To shorten the distance between base stations (BSs) and user terminals, next-generation mobile communications (6G) plans to install large numbers of remote antenna units (RAUs) on traffic signals and street lights and connect these RAUs to base band units (BBUs) on buildings using terahertz (THz) band fronthaul radio lines capable of data rates that exceed 100 Gbit/s. However, when THz band fronthaul wireless circuits are densely deployed in urban areas, the challenge is to maintain line-of-sight (LOS) between RAUs and BBUs and prevent interference between fronthaul wireless links. In this study, the three-dimensional (3D) radiation pattern of a 300-GHz-band high-gain antenna was measured using the near-field-to-far-field (NF-FF) conversion method, and the accuracy was compared with the far-field measurement results. Moreover, an algorithm for automatically deploying a 300-GHz-band wireless fronthaul link is proposed, which can be used to position BBUs in locations where one BBU can be connected to as many RAUs as possible. Propagation simulations for fronthaul wireless links placed by the automatic deployment algorithm, using the measured 3D radiation patterns from high-gain antennas, show no interference between the fronthaul wireless links.

key words: *terahertz antennas, aperture antennas, near-field radiation pattern, radio communication equipment*

1. Introduction

The next-generation mobile communication standard, 6G, is expected to deploy a terahertz (THz) wireless system to achieve > 100 Gbit/s data rates and > 100 times the existing capacity in the next decade [1], [2]. The need for high capacity also leads to ultra-dense networks. It is ideal to communicate as close as possible, and in an unobstructed environment. Therefore, the placement of remote antenna units (RAUs) on street lights and traffic signals has been investigated [2]–[4]. Considering that not all base stations in such an ultra-dense network have access to fiber, a 300-GHz-band fronthaul wireless link connecting the RAUs and base band units (BBUs) is a viable option.

However, it is difficult to deploy many high-density THz wireless links in line-of-sight (LOS) metropolitan areas because skyscrapers obstruct the LOS of fronthaul links. Automatic deployment algorithms for high-density backhaul links have been investigated to connect THz backhaul wireless links to a LOS environment [5]–[7]. However, these automatic deployment algorithms connect RAUs on buildings with LOS environments and to a certain extend have to rely on a few sites connected by fiber. To date, no algorithms have been considered for deploying BBUs on buildings that cover all RAUs installed on street lights or traffic signals.

Another problem with high-density THz fronthaul wireless links is the interference between the fronthaul links [6]. THz-band fixed wireless links for backhaul/fronthaul links use high-gain antennas, such as Cassegrain antennas (CAs) to reach link distances that range from tens to hundreds of meters. A three-dimensional (3D) antenna radiation pattern model is required to evaluate the interference between fixed wireless links. To evaluate interference between wireless links, antenna radiation pattern models, as described in the ITU-R Recommendations, are commonly used. References [5] and [7] employed the mathematical antenna radiation pattern models described in ITU-R Recommendations F. 699 [8] and F. 1245 [9]. However, the available models are limited to 86 GHz [8], [9]. Additionally, few experimental results exist on the far-field (FF) radiation pattern of a high-gain antenna at 300 GHz. Reference [6] employed a 3D radiation pattern generated by rotating a 2D antenna radiation pattern measured in an anechoic chamber for interference evaluation [10]. However, this antenna pattern has not been evaluated in the far-field region. Moreover, the generated radiation pattern is mathematically generated and no measurements of the actual 3D radiation pattern were used. The FF measurement of a high-gain antenna, even at 300 GHz, requires a transmission distance of several tens of meters [10], [11]. However, in previous studies, experiments were conducted in anechoic chambers, where the transmission distance in the measurements was below the FF boundary. Moreover, conventional FF measurements can be used to acquire two-dimensional (2D) radiation patterns, such as the E- and H-planes of the antenna. Near-field (NF) pattern measurement and near-field-to-far-field (NF–FF) conversion are used to measure the three-dimensional radiation patterns of high-gain antennas in an anechoic chamber [12]. It has been reported that NF–FF conversion is beneficial for evaluating the radiation pattern of low-gain antennas at 300 GHz [13], [14]. A few studies have investigated NF–FF conversion for high-gain antennas at 300 GHz, but its effectiveness for high-gain antennas at 300 GHz has not been examined. Tanaka et al. reported the measurement of the NF antenna pattern of a CA at 300 GHz and applied NF–FF conversion [15]. However, the measurement of the gain of a high-gain

Manuscript received October 6, 2023.
Manuscript revised February 13, 2024.
Manuscript publicized June 28, 2024.
[†]Dept. of Information and Communication Systems Engineering, Chiba Institute of Technology, Narashino-shi, 275-0016 Japan.
[††]Technische Universität, Braunschweig, Germany.
a) E-mail: hirata.akihiko@p.chibakoudai.jp
DOI: 10.23919/transcom.2023EBP3162

antenna at 300 GHz has not yet been reported because it requires the measurement of the received power at a distance beyond the FF boundary.

In this study, the 2D FF radiation patterns of a 300-GHz-band CA with a gain of 45 dBi beyond the FF boundary were measured and compared to the 3D radiation pattern obtained by near-field-to-far-field (NF-FF) conversion. The absolute antenna gain of the CA was also measured by converting it into an orthogonal horn antenna, whose absolute antenna gain was measured using the three-antenna method [16]. Finally, the interference between the 300-GHz-band wireless fronthaul links deployed by an automatic planning algorithm using the measured 3D CA radiation pattern was evaluated. No previous reports exist on automatic deployment algorithms for connecting RAUs installed on street lights or traffic signals to BBUs on buildings. Moreover, this is also the first time that measured antenna radiation patterns have been used to evaluate interference between fronthaul links connecting RAUs installed on street lights or traffic signals and BBUs on buildings.

2. Measurement of Far-Field Characteristics

Figure 1 shows the photographs of the antennas whose characteristics were measured. The FF gain and radiation pattern of the CA with a diameter of 150 mm were measured. The waveguide size of this antenna was WR3.4 (220–330 GHz). The typical gain of the CA, calculated by the manufacturer through simulations, was 47 dBi. Moreover, we measured the absolute gains of a waveguide probe antenna and an orthogonal horn antenna, which were used as reference antennas for measuring the absolute antenna gains. The specifications of the antennas are listed in Table 1. The WR3.4 rectangular waveguide (1.16 mm × 0.73 mm) formed the input ports of these antennas.

To evaluate antenna characteristics such as antenna gain and radiation pattern, these antenna characteristics must be measured at a distance beyond the FF boundary. The FF boundary for reflector antennas (R_f) can be expressed as follows:

$$R_f = \frac{2D^2}{\lambda}, \tag{1}$$

where D indicates the diameter of the antenna aperture, and λ signifies the wavelength. If D is 150 mm and λ is 1.02 mm (293.4 GHz), R_f becomes approximately 45 m. Therefore, it is impossible to measure the gain and radiation pattern of the CA listed in Table 1 in an ordinary anechoic chamber. To perform outdoor experiments at 300 GHz, a prototype 300-GHz-band Tx and Rx was constructed. The Tx and Rx configurations are shown in Fig. 2. In the Tx, the 16.3 GHz local oscillator (LO) signal was multiplied by 18 to obtain 293.4 GHz. The output power is 0.8 dBm. In the Rx, a subharmonic mixer downconverts the RF signal to a 24 MHz intermediate frequency (IF) signal, and the received RF power can be calculated by measuring the magnitude of the IF signals.

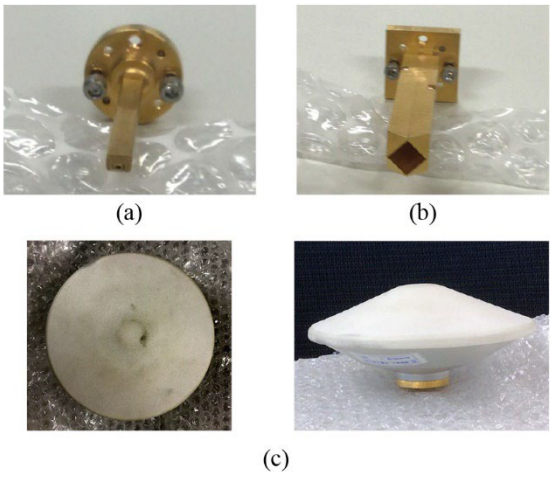


Fig. 1 Photographs of (a) waveguide probe antenna, (b) orthogonal horn antenna, and (c) Cassegrain antenna.

Table 1 Types of antennas used in this study.

Antenna type	Aperture size	Gain (typ.) [1*]
Cassegrain	150 mm (diameter)	47 dBi
Orthogonal horn	5.6 mm × 5.6 mm	25 dBi
Waveguide probe	1.16 mm × 0.73 mm	-

1*: These values are typical specifications provided by the antenna manufacturer and not the measured values.

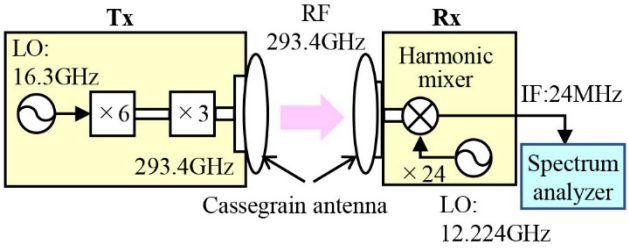


Fig. 2 Schematic diagram of 300-GHz-band Tx and Rx.

We measured the gain of the CA by converting it into an orthogonal horn antenna whose absolute antenna gain was measured using the three-antenna method [16]. In the experiment, we used two orthogonal horn antennas and two waveguide probe antennas, which are listed in Table 1. Moreover, we used a vector network analyzer (VNA) with WR3.4 frequency extenders (220–330 GHz). The distance between the two antennas was 0.5 m. The measured gains of the antennas are shown in Fig. 3. The gains of the orthogonal horn antennas at 293.4 GHz were 25.1 dBi and 25.4 dBi, and those of the waveguide probe antenna were 7.1 and 7.2 dBi. The gain of the CA using the Tx and Rx was evaluated, as shown in Fig. 2. First, we measured the received power of the Rx when both the Tx and Rx used a CA with a transmission distance of 50 m. The height of the CA from the ground was 1.2 m. When the distance between the Tx and the Rx was 50 m, the radius of the first Fresnel zone was 0.11 m. The offset angle of the CA from the ground at the midpoint between Tx

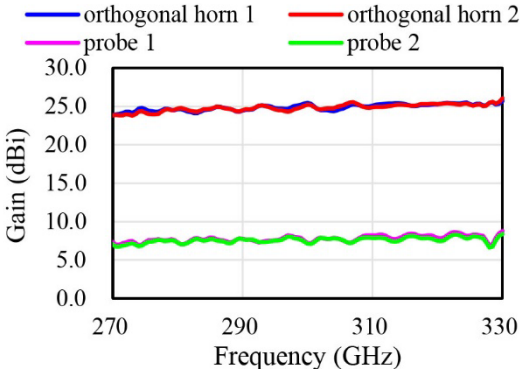


Fig. 3 Gains of orthogonal horn antennas and waveguide probe antennas measured using the three-antenna method.

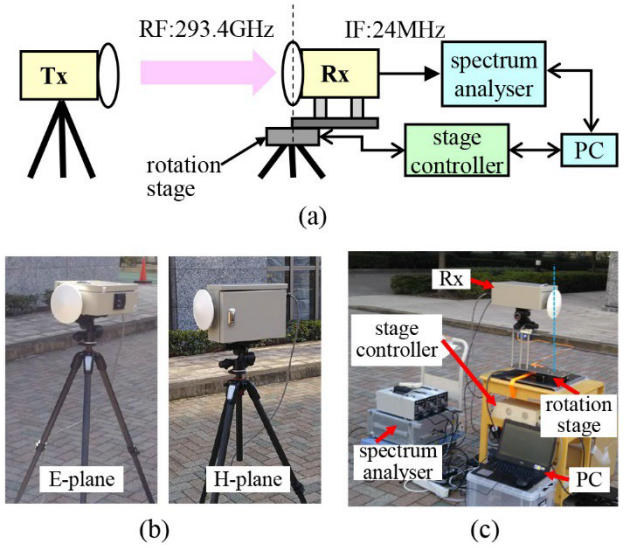


Fig. 4 (a) Experimental setup for the measurement of the FF radiation pattern. (b) Photographs of Tx for the measurement of the E- and H-planes of the antenna radiation patterns. (c) Photograph of Rx.

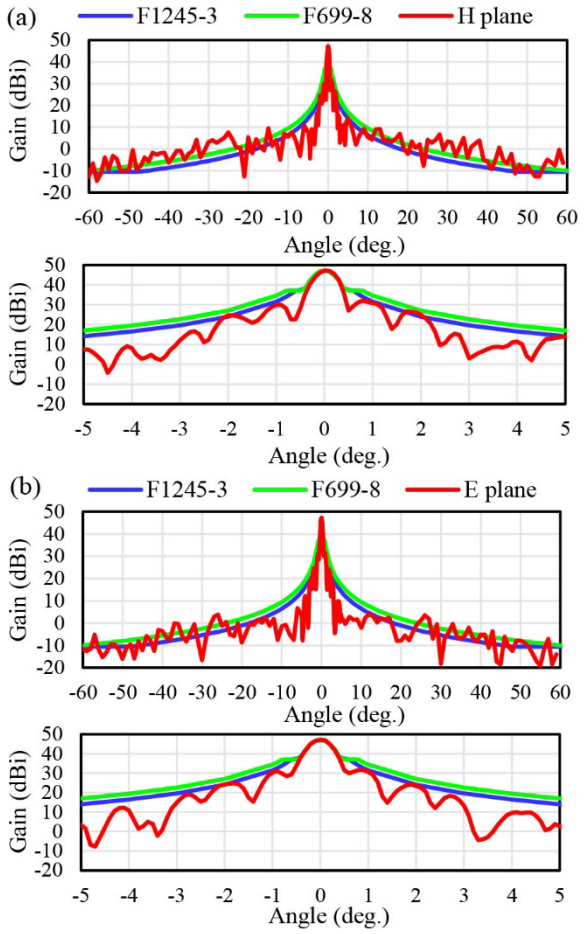


Fig. 5 FF radiation patterns of the CA in (a) H-plane and (b) E-plane. The radiation patterns described in the Recommendations ITU-R F. 699-8 and F. 1245-3 are also shown.

and Rx was 2.7°, and the gain of the antenna at that offset angle was approximately −28 dBc; therefore, the magnitude of the reflected wave was −28 dB smaller than that of the direct wave. These results indicate that at an antenna height of 1.2 m, the reflected waves on the ground have little effect on the measurement results of the antenna characteristics. Subsequently, the received power of the Rx was measured after changing its antenna to an orthogonal horn antenna. The gain of the orthogonal horn antenna at 293.4 GHz was 25.1 dBi. The CA gain was calculated by adding the difference in the received power when using the orthogonal horn antenna. The measured gain of the CA, which was calculated by adding the difference in the received power when using the orthogonal horn antenna was 47.2 dBi.

Subsequently, the radiation patterns of the CA were measured. A diagram and a photograph of the measurement setup are shown in Fig. 4. The Rx was attached to a rotation stage, and the received power and rotation angle were recorded using a PC. The Tx and Rx were mounted on tripods at the bottom and sides of the housing. Therefore, the radiation patterns in the E- and H-planes of the antenna can only be measured when the stage is rotated horizontally. Rx was fixed to the rotating stage so that the center of the antenna coincides with the center of the rotating stage. The rotation was performed in steps of 0.1° from −10 to +10°, and in steps of 1.0° from ±10° to ±60°. The measured radiation pattern of the CA is shown in Fig. 5. The gain of the CA at 0° was set to 47.2 dBi based on the measurement results of the antenna gain. The measured HPBW of the CA was approximately 0.5° in both H- and E-planes, and this value was almost the same as the simulation value provided by the antenna manufacturer. The fluctuation of the gain below 0 dBi (at an offset angle of 10°) originated from the receiver noise level and wind-induced vibrations of the Tx and Rx antennas. The first sidelobe levels in the H- and E-planes were approximately −16.1 dBc and −16.2 dBc, respectively. Mathematical models of the average radiation pattern for a fixed point-to-point wireless system antenna below 86 GHz are provided in Recommendations ITU-R F.699-8 [8] and ITU-R F. 1245-3 [9]. These antenna radiation pattern models are used in interference assessments. Therefore, it is

desirable that the measured gain is less than the gain of the model at all angles. The radiation patterns of Recommendations ITU-R F.699-8 and F.1245-3 are shown in Fig. 5. The measured gain between −10° and +10° was smaller than the gain of the ITU-R Recommendation models. At angles greater than 20°, some of the measured gains exceed those of the ITU-R Recommendation models by 10 dB. These results indicate that the ITU-R Recommendation models should be modified for large offset angles. The gain of the antenna is expressed in Eq. (2) and Eq. (3) for ITU-R F. 699-8 and F. 1245-3, respectively.

$$G(\varphi) = 32 - 25\log\varphi \quad \text{for} \quad 0.59 < \varphi < 48 \tag{2}$$

$$G(\varphi) = 29 - 25\log\varphi \quad \text{for} \quad 0.78 < \varphi < 120 \tag{3}$$

Equations (2) and (3) need to be modified to Eqs. (4) and (5), respectively, to match the results in the graph shown in Fig. 5.

$$G(\varphi) = 42 - 25\log\varphi \quad \text{for} \quad 0.59 < \varphi < 48 \tag{4}$$

$$G(\varphi) = 39 - 25\log\varphi \quad \text{for} \quad 0.78 < \varphi < 120 \tag{5}$$

3. NF-to-FF Measurement

Evaluating the FF characteristics of a high-gain antenna has several requirements, such as a radio station license and a high-sensitivity receiver. Conventional FF measurements can acquire 2D radiation patterns. Alternatively, NF measurements can be performed in an anechoic chamber, and highly stable and high-sensitivity measurement equipment can be used, such as a VNA. For the NF measurement, the amplitude and phase distribution of the field near the antenna surface were measured using a probe antenna, and the FF characteristics were derived using numerical methods [11], [12], [17], [18]. Figure 6(a) and (b) show a schematic diagram and a photograph, respectively, of the experimental setup for measuring the antenna NF pattern. The probe antenna was placed 2 mm from the CA surface and was moved to map the NF distribution. The map area was 150 mm × 150 mm. The amplitude and phase of the 293.4 GHz signal were measured using a VNA. The pitch width of the waveguide probe antenna was 0.5 mm. Antenna axis alignment is a problem when measuring the NF pattern of a large-aperture antenna. The 293.4 GHz signal has a short wavelength (1.02 mm), and a small tilt in the antenna axis affects the phase distribution in the large aperture of the antenna. When the antenna is tilted by 1°, the distance between the measurement plane of the probe antenna and the aperture plane of the antenna changes by a maximum of 2.6 mm. This difference corresponds to a phase difference of 940° for a 293.4 GHz signal.

Figure 7(a) and (b) show the magnitude and phase, respectively, of the NF radiation pattern from the CA, when the antenna axes were visually aligned. The effect of shadowing by the hyperbolic subreflector was observed at the center of the NF magnitude distribution, and the electric field magnitude distribution was circularly symmetric. However, the

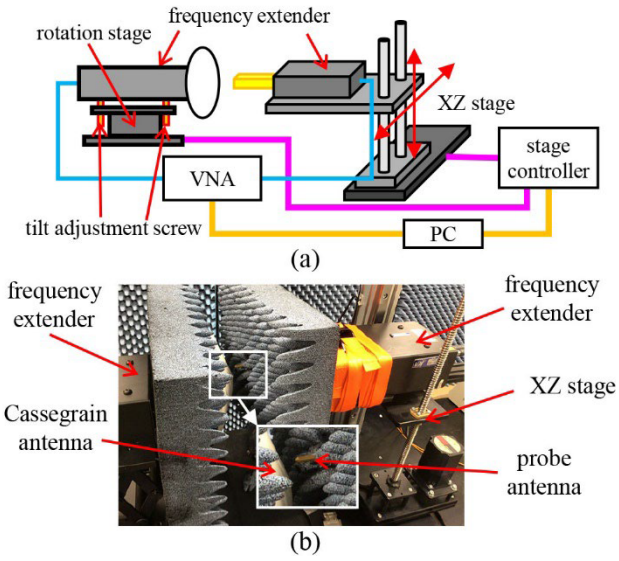


Fig. 6 (a) Schematic and (b) photograph of the experimental setup for the measurement of antenna near-field pattern.

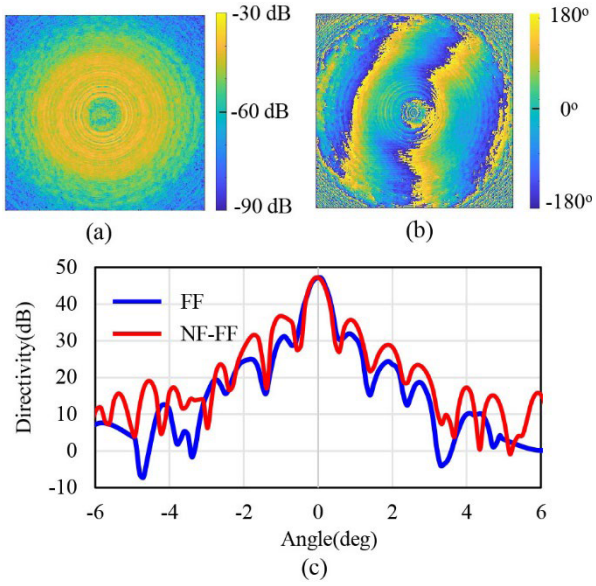


Fig. 7 (a) Measured S_{21} magnitudes and phases (b) of NF pattern of CA before axis alignment. (c) Measured FF radiation pattern and NF (Fig. 7(a) and (b))–FF conversion results of CA in the H-plane.

phase distribution was tilted because the antenna axis was inclined toward the measurement plane. Based on the phase shift in the antenna aperture, the tilt angles of the antenna axis were 3° and 1° along the x and z axes, respectively. Figure 7(c) shows the NF-FF conversion results in the H-plane. The measured FF radiation patterns are also shown. The sidelobe levels of the NF–FF conversion results using the NF patterns shown in Fig. 7(a) and (b) are 8 dB larger than the measured values of the FF pattern, and a sidelobe level divergence exists between the measured FF pattern and the NF-FF conversion results [19].

To align the antenna axis perpendicular to the measure-

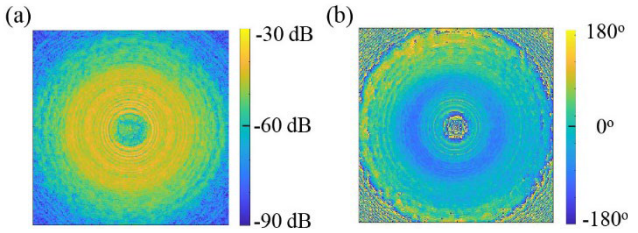


Fig. 8 (a) Measured S_{21} magnitudes and phases (b) of NF pattern of CA after axis alignment.

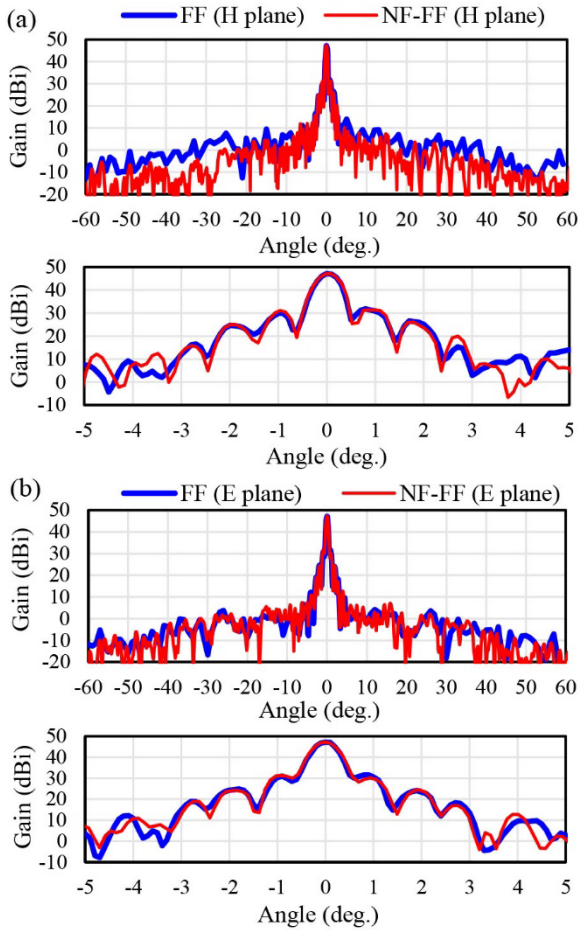


Fig. 9 Measured FF radiation pattern and NF–FF conversion results of LA in (a) H-plane and (b) E-plane.

ment plane, a rotation stage with an angular resolution of 0.01° was used for the frequency extender to which the CA was attached and the antenna axis was adjusted by measuring the phase distribution along the x-axis. Moreover, we adjusted the tilt of the frequency extender along the z-axis using a tilt-adjustment screw at the bottom of the frequency extender. Figure 8(a) and (b) show the magnitude and phase, respectively, of the NF pattern of the CA after the alignment of the antenna axis. The phase of the electric field was circularly symmetric owing to the alignment of the antenna axis. Figure 9 shows the FF patterns calculated from the NF measurement results shown in Fig. 9(a) and (b). Table 2 presents the results of the comparison between the measured

Table 2 Comparison of the measured FF characteristics and NF–FF conversion results of CA at 293.4 GHz.

		H-plane		E-plane	
		FF	NF–FF	FF	NF–FF
HPBW (degree)		0.5	0.50	0.5	0.50
+1st side lobe	angle	0.8	0.76	0.9	1.03
	level	31.9	31.3	31.6	29.7
−1st side lobe	angle	-0.9	-0.84	-0.9	-0.92
	level	31.1	29.9	31.0	31.4
+2nd side lobe	angle	1.7	1.76	1.9	1.95
	level	26.6	26.1	24.2	24.4
−2nd side lobe	angle	-2.0	-1.99	-1.8	-1.83
	level	24.6	25.0	24.8	24.2
+3rd side lobe	angle	2.7	2.70	2.8	2.75
	level	15.3	20.0	18.0	17.0
−3rd side lobe	angle	-2.7	-2.79	-2.8	-2.75
	level	16.6	15.70	18.9	19.00

FF characteristics and the NF–FF conversion results of the CA at 293.4 GHz. The difference in sidelobe angle between the FF and NF–FF was less than 0.1°, and the difference in sidelobe level between them was less than 1.1 dB except for the +1st and +3rd sidelobes. These results indicate that the measured FF pattern and NF–FF conversion results from the main lobe to 3rd sidelobe were in good agreement. When the offset angle was less than 10°, the sidelobe levels of the FF pattern and the NF–FF conversion results were in good agreement. However, at an offset angle of more than 10° in the H-plane, the sidelobe level of the FF pattern was 5–10 dB higher than that of the NF–FF conversion results.

4. Interference Evaluation of Automatically Deployed Fronthaul Links Using the Measured NF-FF Radiation Pattern

In 6G mobile wireless fronthaul, it is planned to mount 300-GHz-band RAUs on traffic signals and street lights to achieve a short transmission distance in LOS environments, and THz fronthaul wireless links are being considered for connecting these RAUs. However, it is difficult to deploy many high-density THz wireless links in metropolitan areas with LOS environments and without interference between fronthaul links. We propose an algorithm for automatic deployment of a 300-GHz-band wireless fronthaul link. In this algorithm, the BBUs are installed on the rooftops of buildings in such a manner that they can be connected to the LOS environment with many RAUs installed on traffic signals or street lights.

Judgement of the LOS environment between Tx and Rx was performed as follows. First, we determined the 3D coordinates of the installed Tx and Rx using the Shinjuku building data. Subsequently, Tx and Rx were connected using 3D line segments whose 3D coordinates were calculated using the 3D coordinates of Tx and Rx. We then calculated the XY coordinates of the points where the computed line segments and the building polygons intersect in the XY plane. If the Z coordinate in the XY coordinate of both ends of the intersected line segment is higher than the height of

the building in the same XY coordinate, it is judged to have a clear LOS.

In addition, BBUs should be placed in locations where one BBU can be connected to as many RAUs as possible to reduce the number of expensive BBUs. In densely arranged fronthaul wireless links, interference between the wireless links occurs due to antenna side lobes. To evaluate the interference power between fronthaul wireless links, the actual 3D radiation pattern data of the antennas are required. Therefore, a radio wave propagation simulation was conducted using the 3D radiation pattern of a Cassegrain antenna measured by the NF-FF conversion shown in Sect. 3, and the cumulative distribution function (CDF) of the signal-to-interference-noise ratio (SINR) was calculated in densely arranged fronthaul wireless links. The automatic deployment algorithm for fronthaul wireless links entails the following seven steps.

Step 1: Based on Street View information, 82 RAUs were placed at the actual positions of traffic signals and street lights in a 750 m × 750 m area in the Shinjuku area of Tokyo. All RAUs are listed on the list of unconnected RAUs.

Step 2: The candidate points were placed 5 m above the roof of the building at 1 m intervals along the edge of the building.

Step 3: For all candidate points identified in Step 2, the number of RAUs that are on the unconnected list were counted, and they were less than 100 m away from the candidate point, and can be connected to the candidate point with the LOS environment.

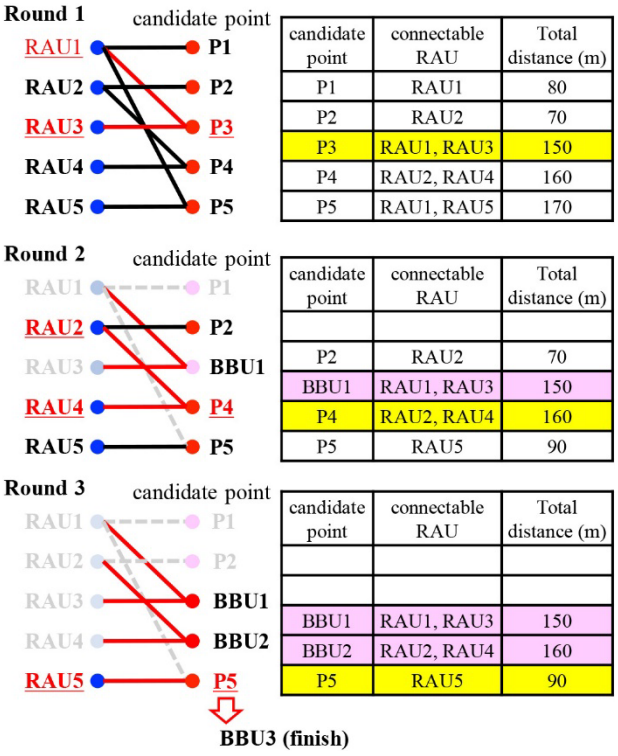
Step 4: The candidate points with the highest number of connectable RAUs were extracted.

Step 5: Among the extracted candidate points, the candidate point with the smallest total transmission distance was assumed to be the position of the BBU by adding all distances to the connectable RAUs. The installed BBUs and RAUs that can be connected to the BBU are listed in the list of BBU-RAU pairs.

Step 6: RAUs that can be connected to the BBUs installed in Step 5 were deleted from the list of unconnected RAUs.

Step 7: Returned to Step 3 and continued the loop of Steps 3–6 until the list of unconnected RAUs became empty.

Figure 10 shows an example of a BBU-RAU pair based on the automatic deployment algorithm. In this example, there were five RAUs and five candidate points. In round 1, lines are drawn between candidate points and RAUs that have a clear LOS. Of the five candidate points, P3, P4, and P5 had a LOS between the two RAUs. The third column of the table in Fig. 10 shows the total distances between the candidate point and the RAUs with LOS. Because P3 had the shortest total distance among P3, P4, and P5, P3 was selected as BBU1. In round 2, RAU1 and RAU3 were connected to BBU1 and P1 was connected to RAU1, and RAU1, RAU3, P1 are removed from the list. Among the candidate points P2, P4, and P5 remaining on the list, the candidate point with the most RAUs was P4. P4 had an LOS with two RAUs (RAU2 and RAU4), and P4 was selected as BBU2. In round 3, RAU2 and RAU4 were connected to BBU2 and P2 was



candidate point	connectable RAU	Total distance (m)
P1	RAU1	80
P2	RAU2	70
P3	RAU1, RAU3	150
P4	RAU2, RAU4	160
P5	RAU1, RAU5	170

candidate point	connectable RAU	Total distance (m)
P2	RAU2	70
BBU1	RAU1, RAU3	150
P4	RAU2, RAU4	160
P5	RAU5	90

candidate point	connectable RAU	Total distance (m)
BBU1	RAU1, RAU3	150
BBU2	RAU2, RAU4	160
P5	RAU5	90

Fig. 10 Example for making the BBU-RAU pair based on the automatic deployment algorithm.

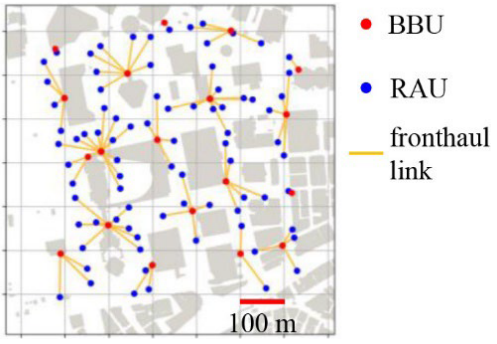


Fig. 11 Example of the automatically planned 300-GHz-band wireless fronthaul links in Shinjuku.

connected to RAU2 and RAU2, RAU4, and P2 were removed from the list. The remaining RAU in round 3 was RAU5. Because the candidate point connected to RAU5 was P5, P5 was selected as BBU3. As a result, all RAUs were connected to the BBU, and the BBU selection loop was concluded.

Figure 11 shows an example of automatically planned 300-GHz-band wireless fronthaul links in Shinjuku. In this algorithm, the Shinjuku building model was used in a 750 m × 750 m area. Among the candidate points, 19 BBUs were selected to be connected with all 82 RAUs placed at the actual traffic signals and street light positions. Next, the signal-to-noise ratio (SNR) and signal-to-interference noise power (SINR) of the 300-GHz-band wireless backhaul links was simulated using a radio-wave propagation simulator based

Table 3 Specification of the 300-GHz-band wireless backhaul link used in the simulation.

Parameter	Value	Remarks
Tx power [dBm]	5	
Data rate [Gbps]	BW/1.2	due to 20% roll-off,
Modulation	64-QAM	
NF [dB]	10	T=300 K
Antenna Gain [dBi]	50	both TX and RX
Payload Rate	0.9	both TX and RX

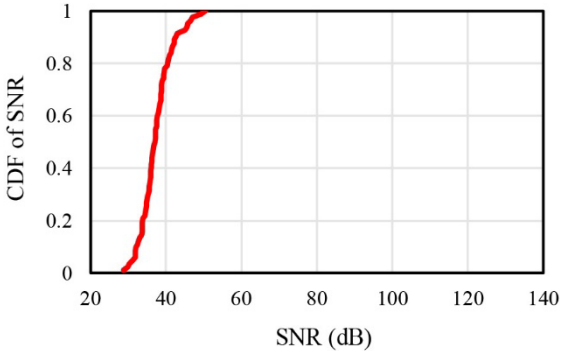


Fig. 12 Simulation results of CDF of SNR for the 300-GHz-band fronthaul wireless links shown in Fig. 11.

on the ray-tracing method (Remcom, Wireless Insite). The specifications of the 300-GHz-band wireless backhaul link used in the simulation were determined based on the target specification of the 300-GHz-band wireless backhaul links in the ThoR project (Table 3) [20]. The EU-Japan joint project "ThoR" was working on the development of 300-GHz-band wireless backhaul links that can achieve a data rate of over-100-Gbit/s [1]. For the 300-GHz-band wireless backhaul link system presented in Table 3, the standard SNR requirement is 22 dB for 100-Gbit/s data transmission [20]. To evaluate the interference between three-dimensionally arranged wireless links, a 3D radiation pattern of the antennas is required. Far-field measurements of antennas can only provide 2D radiation patterns. Therefore, we used the 3D radiation pattern of a 300-GHz-band Cassegrain antenna obtained by the NF-FF conversion shown in Sect. 3.

Figure 12 shows the simulation results for the CDF of the SNR. These results indicate that the SNR of all fronthaul wireless links exceeds the required SNR (22-DB) for 100-Gbit/s data transmission. The CDF of the SNR below 40 dB was 78%, and the maximum SNR was 50.4 dB. Figure 13 shows the simulation results for the CDF of the SINR. When Txs are set at BBUs on buildings and Rxs are set at RAUs on the ground (Model A), the interference power is expected to be low because the RAUs are several tens of meters apart. However, when Txs are at RAUs on the ground and Rxs are at BBUs on buildings (Model B), multiple Rxs are placed at the same location, and multiple Txs radiate THz waves toward the Rxs, and this arrangement may cause interference between the radio links. Therefore, we simulated the SINR of the two models (Model A and B). Figure 13 indicates that the SINR of all fronthaul wireless link exceeds the required

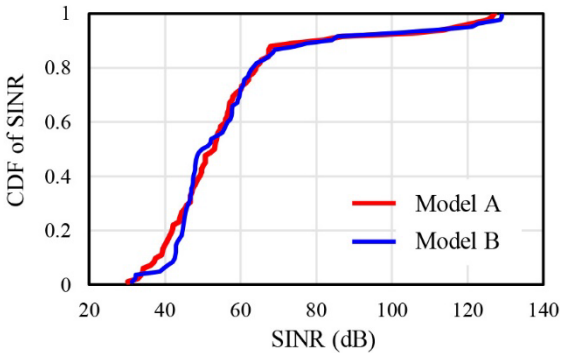


Fig. 13 Simulation results of CDF of SIR for the 300-GHz-band fronthaul wireless links shown in Fig. 11.

SNR (22-DB) for 100-Gbit/s data transmission. The CDF of the SNR below 40 dB were 15% and 6% for Models A and B, respectively, and the maximum SINR were 127 dB and 129 dB for Models A and B, respectively. These results indicate that the interference power between the 300-GHz-band fronthaul wireless links arranged by the automatic deployment algorithm does not affect the transmission speed owing to the use of a high-directivity antenna.

5. Conclusion

We evaluated the interference between 300-GHz-Band fronthaul links arranged by an automatic deployment algorithm with a radio propagation simulator using the measured 3D radiation pattern of a high gain Cassegrain antenna with a gain of 47.2 dBi. The 3D radiation pattern of a 300-GHz-band high-gain antenna was measured using the NF-FF conversion method and the accuracy was compared with the 2D radiation pattern obtained by performing outdoor FF measurements. The results indicated that the NF–FF conversion is a valid method for evaluating the main lobe and sidelobes close to the main lobe of a 300-GHz-band high-gain antenna. An algorithm for the automatic deployment of a 300-GHz-band wireless fronthaul link that can place BBUs in locations where one BBU can be connected to as many RAUs as possible was also proposed. We succeeded in connecting 82 RAUs on street lights and traffic signals to 19 BBUs installed on the rooftops of buildings with an LOS environment. The interference between fronthaul links arranged with an automatic deployment algorithm using a radio propagation simulator that integrates the measurement results of the 3D radiation pattern of the 47.2 dBi Cassegrain antenna was evaluated. The simulation results indicated that the SINR of all fronthaul links exceeded 22 dB, which is required for 100-Gbit/s data transmission.

Acknowledgments

These results were obtained from research (Nos. 19601, 00401, and 04301) commissioned by the National Institute of Information and Communications Technology (NICT), Japan.

References

[1] T. Kürner, A. Hirata, B.K. Jung, E. Sasaki, P. Jurcik, and T. Kawanishi, "Towards propagation and channel models for the simulation and planning of 300 GHz backhaul/fronthaul links," 2020 XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science, Rome, Italy, pp.1–4, 2020. DOI: 10.23919/URSIGASS49373.2020.9232186.
[2] White_PaperEN_v4.0.pdf https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technolology/whitepaper_6g/DOCOMO_6G_
[3] ThoR. (2018 Nov.). Deliverable D2.1: Requirements for B5G backhaul/fronthaul. [Online]; https://thorproject.eu/results/deliverables/#overall-system-concept
[4] A. Hirata, B.K. Jung, and P. Jurcik, "Backhaul/fronthaul outdoor links," THz Communications — Paving the Way Towards Wireless Tbps, T. Kürner, D. Mittleman, and T. Nagatsuma, eds., pp.117–122, Springer, 2022
[5] B.K. Jung and T. Kürner, "Automatic planning algorithm of 300 GHz backhaul links using ring topology," 2021 15th European Conference on Antennas and Propagation (EuCAP), Dusseldorf, Germany, pp.1–5, 2021. DOI: 10.23919/EuCAP51087.2021.9411010.
[6] R. Okumura and A. Hirata, "Automatic planning of 300-GHz-band wireless backhaul link deployment in metropolitan area," Proc. 2020 International Symposium on Antennas and Propagation (ISAP), Osaka, Japan, pp.541–542, 2021. DOI: 10.23919/ISAP47053.2021.9391385.
[7] B.K. Jung, T. Kürner, "Automatic planning algorithms for 300 GHz wireless backhaul links," IEICE Trans. Commun., vol.E105-B, no.6, June 2022. DOI: 10.1587/transcom.2021ISI0002
[8] https://www.itu.int/rec/R-REC-F.699/en
[9] https://www.itu.int/rec/R-REC-F.1245/en
[10] H. Sawada, A. Kanno, N. Yamamoto, K. Fujii, A. Kasamatsu, K. Ishizu, F. Kojima, H. Ogawa, and I. Hosako, "High gain antenna characteristics for 300 GHz band fixed wireless communication systems," 2017 Progress in Electromagnetics Research Symposium - Fall (PIERS - FALL), Singapore, pp.1409–1412, 2017. DOI: 10.1109/PIERS-FALL.2017.8293350.
[11] A. Hirata, K. Fujii, N. Sekine, I. Watanabe, and A. Kasamatsu, "Characterisation of terahertz antenna for beyond 5G systems," International Workshop on Photonics Applied to Electromagnetic Measurements, pp.1–2, Nov. 2019.
[12] C.A. Balanis, Antenna Theory, Analysis and Design, 2nd ed., Wiley, 1997.
[13] Y. Tanaka, G. Ducournau, C. Belem Goncalves, F. Gianesello, C. Luxey, I. Watanabe, A. Hirata, N. Sekine, A. Kasamatsu, and S. Hisatake, "Photonics-based near-field measurement and far-field characterization for 300-GHz band antenna testing," IEEE Open J. Antennas Propag., vol.3, pp.24–31, 2022. DOI: 10.1109/OJAP.2021.3133470.
[14] S. Hisatake, "Near-field measurement and far-field characterization of antennas in microwave, millimeter-wave and THz wave band based on photonics," 2022 International Symposium on Antennas and Propagation (ISAP), Sydney, Australia, pp.351–352, 2022. DOI: 10.1109/ISAP53582.2022.9998761.
[15] Y. Tanaka, H. Arisesa, A. Kanno, N. Sekine, J. Nakajima, and S. Hisatake, "Near-field measurement and far-field characterization of a high-gain Cassegrain antenna at 300 GHz band base on a photonics technology," Proc. 2020 XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science, Rome, Italy, 2020, pp.1–3, 2020.
[16] M. Sakasai, H. Masuzawa, K. Fujii, A. Suzuki, K. Koike, and Y. Yamanaka, "Evaluation of uncertainty of horn antenna calibration with the frequency range of 1 GHz to 18 GHz," J. National Institute of Information and Communications Technology, vol.53, no.1, pp.30–42, 2006.
[17] D. Paris, W. Leach, and E. Joy, "Basic theory of probe-compensated near-field measurements," IEEE Trans. Antennas Propag., vol.26, no.3, pp.373–379, 1978.
[18] M.M. Leibfritz and P.M. Landstorfer, "Full probe-correction for nearfield antenna measurements," Proc. 2006 IEEE Antennas and Propagation Society International Symposium, Albuquerque, NM, USA, pp.437–440, 2006.
[19] K. Watanabe, A. Hirata, I. Watanabe, N. Sekine, and A. Kasamatsu, "Measurement of far field radiation pattern of 300 GHz-band Cassegrain antenna," Proc. 2021 International Symposium on Antennas and Propagation (ISAP), Taipei, Taiwan, pp.1–2, 2021. DOI: 10.23919/ISAP47258.2021.9614528.
[20] https://thorproject.eu/wp-content/uploads/2019/07/ThoR_SIKLU_190417_F_WP2-D2.2-Overall-System-Design.pdf

Ken Watanabe received his B.S. and M.S. degrees from Chiba Institute of Technology, in 2021, and 2023, respectively. He is now working at Hitachi Kokusai Electric Inc.

Ryo Okumura received his B.S. and M.S. degrees from Chiba Institute of Technology, in 2020, and 2022, respectively. He is now working at SoftBank Corp.

Akihiko Hirata received his B.S. and M.S. degrees in chemistry, and his Ph.D. Eng. degree in electrical and electronics engineering from Tokyo University, Tokyo, Japan, in 1992, 1994, and 2007, respectively. He joined the Atsugi Electrical Communications Laboratories of Nippon Telegraph and Telephone Corporation (now NTT Device Technology Laboratories) in Kanagawa, Japan, in 1994. He was a senior research engineer and supervisor at NTT Device Technology Laboratories. Since 2016, he has been a professor at the Chiba Institute of Technology. His current research includes millimeter-wave antennas and ultra-broadband millimeter-wave wireless systems. Prof. Hirata is a senior member of IEEE and a senior member of IEICE.

Thomas Kürner received his Dipl.-Ing. degree in Electrical Engineering in 1990, and his Dr.-Ing. degree in 1993, both from University of Karlsruhe (Germany). From 1990 to 1994 he worked with the Institut für Höchstfrequenztechnik und Elektronik (IHE) at the University of Karlsruhe working on wave propagation modelling, radio channel characterization and radio network planning. From 1994 to 2003, he worked in the radio network planning department at the headquarters of the cellular operator E-Plus Mobilfunk GmbH & Co KG, Düsseldorf, where he was team manager radio network planning support responsible for radio network planning tools, algorithms, processes, and parameters from 1999 to 2003. Since 2003, he has been a Full University Professor for Mobile Radio Systems at the Technische Universität Braunschweig. He is currently the chair of the IEEE 802.15 Standing Committee THz. He was the project coordinator of the H2020-EU-Japan project ThoR ("TeraHertz end-to-end wireless systems supporting ultra-high data Rate applications") and Coordinator of the German DFG-Research Unit FOR 2863 Meteracom ("Metrology for THz Communications"). Thomas Kürner is a Fellow of IEEE.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

Strategies for DOA-DNN Estimation Accuracy Improvement at Low and High SNRs*

Daniel Akira ANDO[†a)], ***Student Member***, **Toshihiko NISHIMURA**[†], ***Senior Member***, **Takanori SATO**[†], ***Member***, **Takeo OHGANE**[†b)], **Yasutaka OGAWA**[†], ***Fellows***, ***and*** **Junichiro HAGIWARA**[††], ***Member***

SUMMARY Implementation of several wireless applications such as radar systems and source localization is possible with direction of arrival (DOA) estimation, an array signal processing technique. In the past, we proposed a DOA estimation method using deep neural networks (DNNs), which presented very good performance compared to the traditional root multiple signal classification (root-MUSIC) algorithm when the number of radio wave sources is two. However, once three radio wave sources are considered, the performance of that proposed DNN decays especially at low and high signal-to-noise ratios (SNRs). In this paper, mainly focusing on the case of three sources, we present two additional strategies based on our previous method and capable of dealing with each SNR region. The first, which supports DOA estimation at low SNRs, is a scheme that makes use of principal component analysis (PCA). By representing the DNN input data in a lower dimension with PCA, it is believed that the noise corrupting the data is greatly reduced, which leads to improved performance at such SNRs. The second, which supports DOA estimation at high SNRs, is a scheme where several DNNs specialized in radio waves with close DOA are accordingly selected to produce a more reliable angular spectrum grid in such circumstances. Finally, in order to merge both ideas together, we use our previously proposed SNR estimation technique, with which appropriate selection between the two schemes mentioned above is performed. We have verified the superiority of our methods over root-MUSIC and our previous technique through computer simulation when the number of sources is three. In addition, brief discussion on the performance of these proposed methods for the case of higher number of sources is also given.

key words: *antenna array, DOA estimation, deep neural network, principal component analysis*

1. Introduction

Direction of arrival (DOA) estimation is a very known array signal processing that is extremely important for many wireless applications. One of the most traditional techniques for DOA estimation is the super-resolution multiple signal classification MUSIC/root-MUSIC algorithm [1], [2]. However, this algorithm being classified as spectral-based, it requires the spectral decomposition of the correlation matrix of the antenna array received signal, which makes its online use prohibitive as the array dimension increases. Therefore, investigation of new approaches for DOA estimation, such as deep learning, is a trending research topic.

Deep learning applied to wireless communication problems is receiving much attention from the industry and academia, since performance of such data-driven techniques can greatly surpass traditional model-based techniques [3]. Although offline training of deep neural networks (DNNs) can be computationally costly, once training is finalized, DNNs can be easily deployed online to the specific situation for which they were trained. The complexity of such online implementation of DNNs is also thought to be comparatively light due to the fact that most of this computation relies on matrix multiplication. In fact, several studies, such as [4]–[7], have reported great results from the implementation of deep learning in DOA estimation. In [4], a framework for end-to-end channel and DOA estimation in the context of massive multiple-input multiple-output (massive MIMO) is proposed. In [5], a combination of a detection and DOA estimation network, which reduces the training-set size and makes it possible to train several DNNs corresponding to different position sectors, is presented. In [6], a low-complexity DOA estimation technique for hybrid MIMO systems with uniform circular array at a base station is presented. In [7], a DOA estimation system which is robust to array imperfections is explained. Our research group also tackled this problem in [8]–[11], where we have demonstrated great DOA estimation performance.

Principle component analysis (PCA) is an algorithm used to represent the information contained in a higher dimensional data in a lower dimensional space while keeping intact as much of this information as possible. It is heavily used in areas such as data compression, image analysis, visualization, pattern recognition, regressions, etc. It has been verified that PCA is a very effective technique in order to enhance the performance of machine learning models at the same time that it reduces the number of features in the data, which simplifies these models greatly [12]–[15]. Yet, most studies take advantage of PCA in areas such as image classification, such as [13]. In the DOA estimation field, PCA was also used in [14], [15]. In [14], a PCA-like unsupervised neural network is used to reduce the dimensionality of the training dataset generated from a broadband acoustic signal emitted by a low-altitude and high-subsonic flight target. The authors verified that the performance of their 2-dimensional DOA estimation technique surpasses that of root-MUSIC at

Manuscript received December 26, 2023.
Manuscript revised May 9, 2024.
Manuscript publicized July 18, 2024.

[†]Graduate School/Faculty of Information Science & Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
[††]Faculty of Social Informatics, Mukogawa Women's University, Nishinomiya-shi, 663-8137 Japan.

*A part of this paper was presented at International Symposium on Communications and Information Technologies (ISCIT 2023) [19].

a) E-mail: dakiraando@m-icl.ist.hokudai.ac.jp
b) E-mail: ohgane@m-icl.ist.hokudai.ac.jp

DOI: 10.23919/transcom.2023EBP3217

lower SNRs. In [15], a 1-dimensional narrowband DOA estimation with K-nearest neighbors algorithm is proposed, where PCA is applied to the training dataset in order to reduce the computational complexity of this machine learning algorithm and to remove noise from the signal data. The authors also verified that the performance of their method greatly surpasses root-MUSIC at lower SNRs.

The aim of this paper is to improve the DNN's DOA estimation accuracy, and our contributions are:

- Proposing a method to improve DOA estimation performance at low SNR, which consists of application of PCA to the DNN training, validation and test datasets;
- Proposing a method to improve DOA estimation performance at high SNR, which consists of several separately trained DNNs specialized in radio waves with close DOA;
- Realizing a system binding together the two methods above in order to develop a full technique for DOA estimation at any SNR.

In terms of the first method with PCA, previous studies [14], [15] have not thoroughly evaluated the effects of applying PCA specifically to a DNN input dataset. Therefore, we also give here an extensive and detailed explanation on the effects of PCA at different simulation settings, such as varying a) sizes of the antenna array, b) number of principal components chosen as the new dimension of the input dataset, and c) test SNRs. Note that, in this study, we consider Probabilistic PCA. This is a consequence of the Scikit-learn framework [16], implemented during our numerical simulations, which is based on it. In addition, use of the abbreviation "PCA" will be maintained to keep the notation uncluttered. The second method is based on the observation that, when 3 radio sources are considered, many incorrect estimation cases at higher SNRs are due to radio waves closely impinging onto the antenna array within 20°. Consequently, we offline train different DNNs that are each specialized in close waves impinging at specific regions of the angular spectrum. These new DNNs are then used instead of the conventional DNN, which are expected to produce a more reliable narrow DOA spectrum grid for subsequent DOA detection. Lastly, we use one of our previous strategies for SNR estimation [11] to merge these two proposed methods into a full system capable of operating online while greatly surpassing the performance of our previous technique [9] and root-MUSIC.

Our main study goal in this paper is when there are 3 radio wave sources. As stated in [17], 3 flying objects in the field of view of an antenna array is a possible scenario in air-to-air emitter location or radar systems. Moreover, as verified in [18], at sub-terahertz and line-of-sight indoor office environments of 140 GHz, the average number of subpaths (or multipaths) is significantly small, e.g. mostly ranging between 2 and 5. Therefore, our 3 sources consideration is not only realistic in airborne radar applications based on [17], but also it is a first step towards the goal of radio propagation measurements at sub-terahertz bands [18]. Moreover, it was concluded in [11] that we needed to solve the above mentioned issues (i.e. poor estimation performance at lower and higher SNRs), which arise when the number of sources is simply raised from 2 to 3. However, we also give here a brief discussion on the performance of the proposed methods for the case of 4 and 5 number of sources.

The remainder of this paper is organized as follows. The antenna array model is explained in Sect. 2. Our previous works [9], [19] are detailed in Sect. 3. Our proposed techniques based on these works are presented in Sect. 4. Then, in Sect. 5, we validate our proposed methods though computer simulations while using our past technique and root-MUSIC as benchmark. Lastly, in Sect. 6 our work is concluded.

2. Antenna Array Model

Let there be K radio wave sources located in the far-field region of a uniform linear array (ULA) consisting of L omnidirectional antennas with no mutual coupling and spaced at half-wavelength. These sources are emitting narrowband waves whose planar wavefronts impinge onto the ULA at angles $\boldsymbol{\theta}[\text{degrees}] = [\theta_1, \ldots, \theta_K]^T$ at least 1° apart, where $[\cdot]^T$ indicates the transpose operator. Then, the baseband received signal $\mathbf{x}(t) \in \mathbb{C}^{L\times 1}$ can be modeled by

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{z}(t), \tag{1}$$

where $\mathbf{s}(t) \in \mathbb{C}^{K\times 1}$ is the vector containing the incident radio waves' complex amplitudes, $\mathbf{z}(t) \in \mathbb{C}^{L\times 1}$ is the additive white Gaussian noise vector following a circular complex Gaussian distribution $\mathbf{z}(t) \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_L)$ with zero mean and variance σ^2, where $\mathbf{I}_L$ represents an L-dimensional identity matrix, and $\mathbf{A}(\boldsymbol{\theta})$ is the mode matrix, which accounts for the relative phase delay corresponding to path length difference of the incident waves on each ULA element and is described as

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & \cdots & 1 \\ e^{-j\pi \sin\theta_1} & \cdots & e^{-j\pi \sin\theta_K} \\ e^{-j\pi 2 \sin\theta_1} & \cdots & e^{-j\pi 2 \sin\theta_K} \\ \vdots & \ddots & \vdots \\ e^{-j\pi(L-1) \sin\theta_1} & \cdots & e^{-j\pi(L-1) \sin\theta_K} \end{bmatrix}. \tag{2}$$

Furthermore, the radio waves are assumed to be uncorrelated and received with equal power normalized to one.

For many DOA estimation techniques, the estimated correlation matrix $\hat{\mathbf{R}}_{xx}$ of the received signal is usually used, where this can be calculated by the equation bellow:

$$\hat{\mathbf{R}}_{xx} = \frac{1}{N_{\text{snap}}} \sum_{n=1}^{N_{\text{snap}}} \mathbf{x}(t_n)\mathbf{x}(t_n)^H, \tag{3}$$

where N_{snap} is the total number of snapshots, $\mathbf{x}(t_n)$ represents the nth snapshot taken from the received signal, and $(\cdot)^H$ is the conjugate transpose operator.

3. Authors' Previous Work

3.1 Input and Output Definitions of DNN

The generation procedure of the DNN datasets is described here. Note that these are the original datasets prior to dimensionality reduction through PCA, where they consist of input $\mathbf{u} = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ and target vectors $\mathbf{t} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$. Here, N is the number of samples, $\mathbf{u}_i \in \mathbb{R}^{D_{\text{in}} \times 1}$ and $\mathbf{t}_i \in \mathbb{R}^{D_{\text{out}} \times 1}$ for $i = 1, \ldots, N$ are the ith samples of the input and target vectors with D_{in} and D_{out} features, respectively.

(a) Input Layer

Due to its Hermitian nature, the estimated correlation matrix $\hat{\mathbf{R}}_{xx}$ can be written as in (4). Then, a vector $\mathbf{u}_i$ proper for being fed as input to the DNN can be generated as follows: first, we arrange the diagonal elements of (4) in the first entries of the input vector; next, we take the real $\Re(\cdot)$ and imaginary $\Im(\cdot)$ parts of each lower triangular element column by column and from left to right, subsequently arranging these in the remaining space of the input vector (See (5)). The upper triangular elements can be ignored due to the fact that they are simply the complex conjugate of the lower triangular elements.

$$\hat{\mathbf{R}}_{xx} = \begin{bmatrix} r_{11} & r_{21}^* & \cdots & r_{L1}^* \\ r_{21} & r_{22} & \cdots & r_{L2}^* \\ \vdots & \vdots & \ddots & \vdots \\ r_{L1} & r_{L2} & \cdots & r_{LL} \end{bmatrix} \tag{4}$$

$$\begin{aligned} \mathbf{u}_i = [&r_{11}, \ldots, r_{LL}, \Re(r_{21}), \Im(r_{21}), \ldots, \Re(r_{L1}), \\ &\Im(r_{L1}), \ldots, \Re(r_{L(L-1)}), \Im(r_{L(L-1)})]^T . \end{aligned} \tag{5}$$

The resultant input vector $\mathbf{u}_i \in \mathbb{R}^{D_{\text{in}} \times 1}$ has $D_{\text{in}} = L^2$ features, where each of them corresponds to each unit of the DNN input layer.

(b) Output Layer

We design the DNN output layer in such a way that the DNN should produce an angular spectrum discretized in angle bins, where each of these covers a portion of the spectrum. Therefore, each unit of the DNN output layer corresponds to each angle bin. In this study, an angle spectrum ranging from $-60°$ to $+60°$ is considered. When this spectrum is discretized in steps of $1°$, the total number of angle bins (and thus the number of features D_{out}) becomes 121.

Since the DNN output units represent the probability of incident radio wave onto the corresponding angle bins, the target vector $\mathbf{t}_i = [t_1, \ldots, t_j, \ldots, t_{D_{\text{out}}}]$ can be generated following (6) below.

$$t_j = \begin{cases} 1 & \text{if wave is incident onto the } j\text{th bin} \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

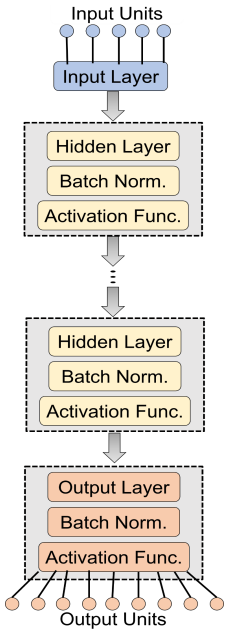


Fig. 1 DNN structure with input, hidden, and output layers. Reprinted from [19] (©2023 IEEE).

where the jth angle bin covers the spectrum region from $j - 61.5°$ to $j - 60.5°$.

3.2 DNN for DOA Estimation

A traditional feed-forward neural network, whose structure consists of an input layer, an output layer and an arbitrary number of hidden layers, is used (Fig. 1). In addition, we insert the batch normalization regularizer [20] in all layers of the DNN in order to improve the overall stability of the learning process. The activation functions for the hidden layers and for the output layer are the rectified linear unit (ReLU) and Sigmoid, respectively. By using Sigmoid as an activation function, we guarantee that the DNN produces output values that can be regarded as probabilities ranging from 0.0 to 1.0. During the learning phase, the DNN weights are updated in accordance to the Adam optimization algorithm [21].

In this work, we investigate mainly two performance metrics: the probability of correct DOA estimation and the root mean squared error (RMSE), where the former is verified during the validation and test phases, and the latter only during the test phase. The probability of correct DOA estimation is calculated as the ratio of the number of correct DOA estimation samples over the total number of evaluated samples, where DOA estimation is only counted as correct when the absolute error of all the DOA estimates $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_K]^T$ within a sample is below a certain estimation tolerance error:

$$\text{Correct DOA} \iff |\theta_j - \hat{\theta}_j| \leq \mu, \forall j \in \{1, \ldots, K\}, \tag{7}$$

where μ is the estimation tolerance, considered to be $0.5°$ here (verification whether the estimated DOA is within the

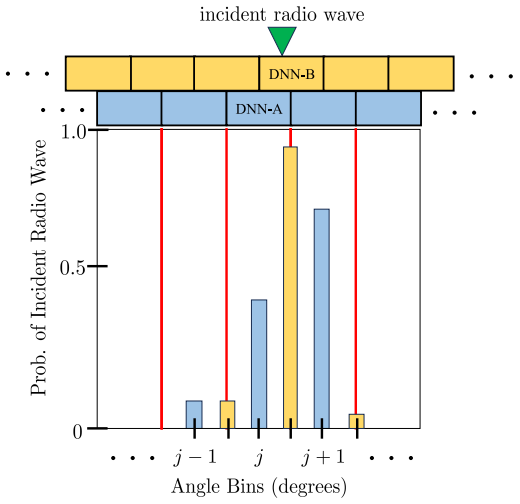


Fig. 2 An example of the Staggered-DNN output, where the radio wave is incident near the right border of the j angle bin of DNN-A. In this example, the $j+1$ bin is mistakenly detected. Cases such as this are one of the verified causes generally leading to incorrect DOA estimation. However, after combining both DNN-A and DNN-B grids, correct DOA detection becomes possible. Reprinted from [19] (©2023 IEEE).

1°-width angle bin). The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{KN_t}\sum_{k=1}^{K}\sum_{n=1}^{N_t}\left(\hat{\theta}_k^{(n)} - \theta_k^{(n)}\right)^2}, \tag{8}$$

where N_t is the total number of test samples. During the validation phase, the DNN weights corresponding to the highest probability of correct DOA estimation are saved for subsequent use at the test phase, where this is done in an effort to avoid overfitting. However, note that the saved weights are not necessarily optimal in terms of RMSE.

Previously we verified that incidence of radio waves onto the vicinity of the angle bin border generally results in incorrect DOA detection due to wrongful excitement of neighboring bins (Fig. 2), thus causing significant decline in overall performance. In [9] we proposed a strategy to cope with such cases. This relies on the training of one additional support DNN (called DNN-B) whose angle grid is stacked up on top of that of the main DNN (DNN-A), where the DNN-B angle grid is shifted by 0.5° with respect to that of DNN-A, thus ranging from −60.5° to +60.5° (totaling 122 angle bins). This strategy was then named Staggered DNN, and we have demonstrated that the spectrum contribution provided with DNN-B enhances the estimation accuracy around the bin borders of the DNN-A bins. Both DNNs are offline trained separately with the same input dataset, but with accordingly modified target datasets reflecting the corresponding angle bin grid. Then, during the test phase, the spectrum grid produced with both DNNs are merged as it is shown in Fig. 2, resulting in a combined angular spectrum grid. Lastly, a DOA detection algorithm is applied on this resultant grid so as to extract the DOA estimates.

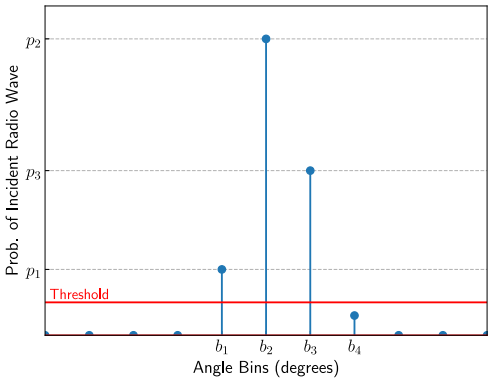


Fig. 3 Illustration of the usage of "Neighbors Weighted Average" on a sample of tested DNN output, assuming only one DOA. For the computation of the DOA estimate, all bins whose probability of incident radio wave is below a certain threshold are ignored (e.g. bin b_4). Reprinted from [19] (©2023 IEEE).

3.3 DOA Detection Algorithm

After calculating the DNN output, or equivalently the angular grid spectrum, it is still necessary to recover the DOA information contained in it. A detection algorithm called "Neighbors Weighted Average" was presented in [19]. Here, we briefly explain it again, and detail the full algorithm in Appendix A.

Various DNN outputs are normally contaminated by spurious bins in the vicinity of those corresponding to true DOA bins. However, by taking advantage of such bins, we managed to develop an algorithm capable of detecting more accurate DOAs than if we had simply chosen the most likely bin by means of, for instance, peak search.

Figure 3 illustrates a DNN output example for the case of only one radio wave. Although no proper optimization procedure has been performed, we have verified that very accurate DOA estimation is possible when the threshold value (straight red line in Fig. 3) is 0.1. Then, the DOA estimate $\hat{\theta}$ can be calculated as:

$$\hat{\theta} = \sum_{i=1}^{3} p_i b_i \Bigg/ \sum_{i=1}^{3} p_i \tag{9}$$

This method has proven to be powerful when there are K clearly distinguished hills of angle bins (for instance, there is only one hill in Fig. 3). On the other hand, the full algorithm described in Appendix is capable of dealing with other cases of less ideal angular spectrum grid.

4. Proposed Strategies for Accuracy Enhancement

4.1 Lower SNRs: Staggered DNN-PCA

The flow chart of the proposed technique for accuracy enhancement at lower SNRs can be seen in Fig. 4. At the training phase, we generate N input samples $\mathbf{u}_i$ for

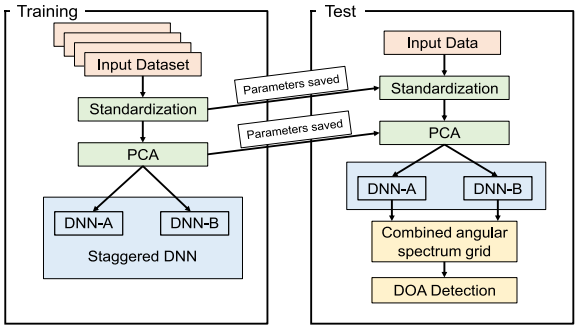


Fig. 4 Flow chart of the proposed Staggered DNN-PCA for lower SNRs, where dimensionality reduction of the DNN input vector is performed with PCA.

$i = 1, \ldots, N$ at 30 dB. Then, these are standardized, resulting in $\bar{\mathbf{u}}_i = \mathbf{\Sigma}^{-1}(\mathbf{u}_i - \boldsymbol{\mu})$. Here, $\boldsymbol{\mu} \in \mathbb{R}^{D_{\text{in}} \times 1}$ and $\mathbf{\Sigma} = \text{diag}\,(\boldsymbol{\sigma}) \in \mathbb{R}^{D_{\text{in}} \times D_{\text{in}}}$ are the vector and diagonal matrix containing the means and standard deviations, respectively, of each feature of the training dataset, where $\boldsymbol{\sigma} \in \mathbb{R}^{D_{\text{in}} \times 1}$ corresponds to the standard deviation vector. By applying PCA to this standardized input dataset we achieve a dimensionality reduction from D_{in} to an arbitrary D_{pca}. In order to derive the PCA parameters (for a more thorough explanation refer to [12]), first the covariance matrix $\mathbf{S}$ of the standardized input dataset must be calculated:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{u}}_i \bar{\mathbf{u}}_i^T. \tag{10}$$

Then, the eigendecomposition of $\mathbf{S}$ is performed:

$$\mathbf{S} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{-1}, \tag{11}$$

where $\mathbf{U} \in \mathbb{R}^{D_{\text{in}} \times D_{\text{in}}}$ and $\mathbf{\Lambda} \in \mathbb{R}^{D_{\text{in}} \times D_{\text{in}}}$ are the matrix containing the eigenvectors of $\mathbf{S}$ and the diagonal matrix containing the corresponding eigenvalues, respectively. After choosing the desired number of dimensions D_{pca} to remain in the new dataset, the parameters $\mathbf{M}$ and $\mathbf{W}$ necessary for dimensionality reduction with PCA can be calculated:

$$\sigma_{\text{pca}}^2 = \frac{1}{D_{\text{in}} - D_{\text{pca}}} \sum_{j=D_{\text{pca}}+1}^{D_{\text{in}}} \lambda_j, \tag{12}$$

$$\mathbf{W} = \mathbf{U}_{\text{pca}} (\mathbf{\Lambda}_{\text{pca}} - \sigma_{\text{pca}}^2 \mathbf{I})^{1/2} \mathbf{R}, \tag{13}$$

$$\mathbf{M} = \sigma_{\text{pca}}^2 \mathbf{I} + \mathbf{W}^T \mathbf{W}, \tag{14}$$

where σ_{pca}^2 can be interpreted as the average variance lost per discarded dimension, $\mathbf{\Lambda}_{\text{pca}} \in \mathbb{R}^{D_{\text{pca}} \times D_{\text{pca}}}$ is the diagonal matrix of the largest D_{pca} eigenvalues λ_j ($j = 1, \ldots, D_{\text{pca}}$), $\mathbf{U}_{\text{pca}} \in \mathbb{R}^{D_{\text{in}} \times D_{\text{pca}}}$ is the matrix with the corresponding eigenvectors (or principal components), $\mathbf{I}$ is the $D_{\text{pca}} \times D_{\text{pca}}$ identity matrix and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix considered to be equal to $\mathbf{I}$ in this study. Finally, the dimension of a sample of the standardized input dataset can be reduced from D_{in} to the chosen D_{pca} by:

$$\mathbf{u}_{\text{pca},i} = \mathbf{M}^{-1} \mathbf{W}^T \bar{\mathbf{u}}_i, \tag{15}$$

where $\mathbf{u}_{\text{pca},i} \in \mathbb{R}^{D_{\text{pca}} \times 1}$ is the representation of $\bar{\mathbf{u}}_i$ in a lower dimension, i.e. the projection points of $\bar{\mathbf{u}}_i$ onto the D_{pca} principal components. This new input dataset is then fed to DNN-A and DNN-B for offline training, while the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\mathbf{M}$ and $\mathbf{W}$ are then saved for later use during the test phase, where new data must go through the same dimensionality reduction process before being fed to the trained Staggered DNN.

Finally, during the test phase, after feeding both DNN-A and DNN-B with a test input sample that went through the PCA process described above, their outputs are combined in order to produce the resultant staggered angular spectrum grid (Sect. 3.2), on which the DOA detection algorithm (Sect. 3.3) is applied.

As it will be shown later in Sect. 5, this strategy provides outstanding accuracy improvement at lower SNRs and number of sources $K = 3$ even when Staggered DNN is trained with a dataset generated at 30 dB. The effectiveness of PCA when K is 4 and 5 is also briefly discussed in Appendix B. As opposed to the noise, which is distributed along all principal components of $\mathbf{S}$, the bulk of the signal information is believed to be mainly distributed along the first D_{pca} principal components. Therefore, improvement in the data SNR is achieved when the $D_{\text{in}} - D_{\text{pca}}$ dimensions are discarded, which ultimately results in more precise DOA estimation. On the other hand, accuracy at higher SNRs is deteriorated due to dimensionality reduction due to loss of signal information, which is only lightly corrupted by noise at such SNRs. Quantitative analysis on the effect of PCA on the input vector SNR, which is believed to be related with estimation performance, is left as our future work.

4.2 Higher SNRs: Staggered Narrow Range DNN

After a preliminary assessment, we observed that waves with close DOA are mostly responsible for incorrect DOA estimation especially at higher SNRs. For instance, when $K = 3$ and the SNR is 30 dB, roughly 80% of all cases of unsuccessful estimation (following (7)) are due to waves lying within a range of 20°. Consequently, in order to overcome this issue especially for the case $K = 3$[†], we have designed a new strategy called "Staggered Narrow Range DNN" (Staggered NRDNN) as shown in Fig. 5.

First, DOA estimation is performed in the exact same way that has been described so far (Fig. 5(a)). After applying Neighbors Weighted Average to the Staggered DNN output (Sect. 3.3), the initial DOA estimates $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_i, \ldots, \hat{\theta}_K\}$ sorted in ascending order are obtained. If these K estimated DOAs are within the range $\Delta\hat{\theta} = \hat{\theta}_K - \hat{\theta}_1 \leq 20°$, then it is very likely that either one was incorrectly detected. At this point a new and more reliable angular spectrum should be produced. To this end, we train 7 different Staggered NRDNNs (7 NRDNN-As and 7 NRDNN-Bs) offline, each covering a

[†]In Appendix B, we discuss the validity of this technique for higher number of sources K.

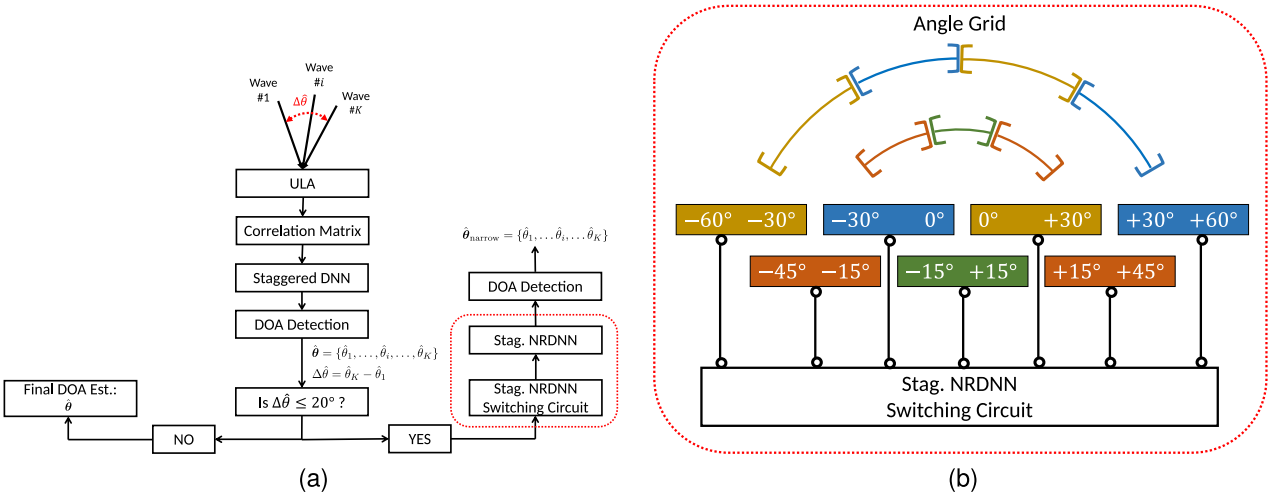


Fig. 5 (a) Flow chart of the proposed Staggered NRDNN strategy for higher SNRs. (b) Visualization of the spectrum grid ranges for which each NRDNN is trained.

predetermined portion of the angle grid (Fig. 5(b)). For instance, there is a Staggered NRDNN covering the angle bins $\{29.5°, 30.0°, 30.5°, \ldots, 60.0°, 60.5°\}$. As it will be shown in Sect. 5, this allows us to create Staggered DNNs specialized in different grid regions, thus making it possible to produce more precise angular spectrum for DOA detection. Next, according to the initial DOA estimates, the appropriate Staggered NRDNN covering them is chosen to produce a new spectrum grid, to which again Neighbors Weighted Average is applied to detect the final DOA estimate $\hat{\boldsymbol{\theta}}_{\text{narrow}}$. On the other hand, if the initial DOA estimates do not lie within the 20° range, then they are kept as the final estimates.

Now we describe the training process for these new NRDNNs. The input and target vector remain the same as in (5) and (6), respectively, where no PCA is involved. In contrast to the main DNN-As and DNN-Bs, where the data was generated by randomly selecting the DOAs $\boldsymbol{\theta}$ from a uniform distribution between $-60.5°$ and $+60.5°$, the training and validation data of the NRDNNs are generated in such a way that all K DOAs $\boldsymbol{\theta}$ lie within a range of $\Delta\theta = \theta_K - \theta_1 \leq 30°$, which in turn is uniformly sampled from $[-60.5°, +60.5°]$. This dataset is then used in all 14 NRDNNs for training and validation. As it will be seen from the simulation results in Sect. 5, good accuracy improvement is achieved even when all NRDNNs are trained with this same dataset. Therefore, there is no need to generate 7 different training datasets corresponding to each specific grid region.

In addition, the precision metric is used during validation in contrast to the probability of correct DOA estimation (Sect. 3.2). The precision of a DNN output is calculated as $n_{\text{TP}}/(n_{\text{TP}} + n_{\text{FP}})$, where n_{TP} and n_{FP} correspond to the number of true positives (DNN angle bin corresponding to a true DOA was excited) and false positives (DNN angle bin where there is no true DOA was excited), respectively. The choice on this metric is due to the impossibility of properly calculating the probability of correct DOA estimation in the considered data generation procedure, where at least 1

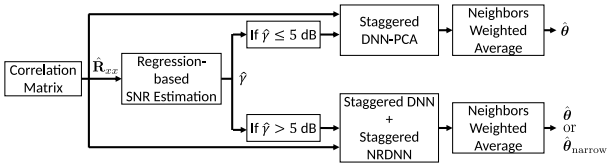


Fig. 6 Flow chart of the proposed full system implemented with Staggered DNN-PCA and Staggered NRDNN with the aid of regression-based SNR estimation [11].

DOA might not lie within the range covered by an NRDNN. Finally, the weights corresponding to the highest precision obtained during the validation of each NRDNN are saved for the test phase.

4.3 Implementation of Full System with the Above Strategies

We have developed two separate strategies for improving the DOA estimation with Staggered DNN at two regions: lower SNRs and higher SNRs. Now it is necessary to bind them together. For this, we use a technique proposed in [11]: Regression-based SNR estimation. The idea is to estimate the SNR $\hat{\gamma}$ from the correlation matrix $\hat{\mathbf{R}}_{xx}$. If $\hat{\gamma} \leq 5\,\text{dB}$, then Staggered DNN-PCA is applied; otherwise, Staggered NRDNN is used (Fig. 6).

In order to estimate the SNR $\hat{\gamma}$, we observed that this in dB and $-\log(\lambda_s)$ are linearly correlated, where λ_s represents the smallest eigenvalue of the correlation matrix $\hat{\mathbf{R}}_{xx}$. Therefore, we obtain a prediction function for $\hat{\gamma}$ with respect to $-\log(\lambda_s)$ by training a regression model based on the ordinary least squares method. Since both our input $(-\log(\lambda_s))$ and output $(\hat{\gamma})$ data are one-dimensional, the linear function that approximates the desired prediction function has only two coefficients: the y-intercept and the slope. These can be calculated by minimizing the residual sum of squares between the observed and the predicted SNRs. After fitting the

training dataset to this model, the following SNR prediction function when $L = 10$ and $K = 3$ is obtained:

$$\hat{\gamma}\,(\mathrm{dB}) = -2.1452 + 9.9923(-\log(\lambda_s)). \tag{16}$$

5. Simulation Results

The performance evaluation of Staggered DNN-PCA (Sect. 4.1), Staggered NRDNN (Sect. 4.2) and the system combining both (Sect. 4.3) is now described here. Training, validation and test datasets must be generated for each Staggered DNN technique and for the regression-based SNR estimation. The parameters for the previously proposed Staggered DNN [9] are shown in Tables 1 and 2. The parameters for the techniques proposed in this paper will be presented in their respective sections. Furthermore, as previously mentioned in Sect. 3, two metrics are used here for performance evaluation: probability of correct DOA estimation and RMSE (the former is calculated in the same fashion as in (7) for root-MUSIC). Their results must be interpreted depending on the type of application. The probability of correct DOA estimation should be of more relevance in such cases where correct DOA estimation of all incoming waves at a time is vital. On the other hand, in such cases where average precision is more important than occasional detection error, RMSE should be mostly considered. Nevertheless, as the RMSE is very sensitive to outliers, other metrics should be regarded concurrently, such as the absolute error median. As mentioned in Sect. 1, we turn our focus in this section on the simulation results for the case of $K = 3$ radio wave sources. For the cases of 4 and 5 sources, a brief discussion on the simulation results is given in Appendix B as a preliminary evaluation.

Table 1 Parameters for Staggered DNN [9] data generation.

Number of antenna elements L	5–15
Number of incident radio waves K	3
Direction of arrival angles θ	random from $-60.5°$ to $+60.5°$
Training and validation SNR	30 dB
Test SNR	0–30 dB
Number of snapshots N_{snap}	100
Number of training data samples	250,000
Number of validation data samples	10,000
Number of test data samples N_t	150,000

Table 2 Parameters for Staggered DNN [9] training.

Input layer units D_{in}	L^2
Hidden layers	5
Units per hidden layer	182
Output layer units D_{out}	DNN-A: 121 DNN-B: 122
Hidden layers activation function	ReLU
Output layer activation function	Sigmoid
Loss function	Binary cross-entropy
Optimizer	Adam
Batch size	256
Number of epochs	500

5.1 Staggered DNN-PCA

All parameters for Staggered DNN-PCA are kept the same as in Tables 1 and 2, except the number of input layer units, which is equivalent to the number of principal components D_{pca} chosen during the dimensionality reduction process described in Sect. 4.1.

First, in Figs. 7 and 8 we show the probability of correct DOA estimation and RMSE of the proposed Staggered DNN-PCA, respectively, with respect to the number of principal components when the number of antenna elements L is varied from 10 to 15 and when the test dataset was generated at 0, 5, 10 and 20 dB. We compare the performance of the proposed technique with that of root-MUSIC, which is shown as horizontal black straight lines in the said figures.

In Fig. 7, we can promptly see that there is an optimal number of principal components especially with respect to 0 dB, and that this optimal value is different for each L. Moreover, when the SNR is 0 dB and $L \geq 11$, the performance of Staggered DNN-PCA surpasses that of root-MUSIC for certain numbers of principal components; in fact, when $L \geq 12$, an improvement of roughly 10% is achieved in terms of the optimal number of principal components. From the blue and orange curves corresponding to 0 and 5 dB, respectively, we can conclude that estimation precision improvement is possible by reducing the size of the DNN input vector with PCA. We believe that, by applying PCA and only selecting the dimensions corresponding to the largest D_{pca} eigenvalues of the covariance matrix $\mathbf{S}$ (refer to (11)), we manage to strongly reduce the noise corrupting the input vector. On the other hand, when the SNR is 20 dB, not only no visible effect from PCA can be seen, but also the proposed method does not surpass root-MUSIC performance. It is believed that information on the signal only lightly corrupted by noise is lost by applying PCA. Therefore, Staggered DNN-PCA appears to be inefficient at higher SNRs (fact that will be verified later in this section).

In Fig. 8, we can see that the choice on the number of principal components mainly affects the RMSE when the SNR is 0 and 5 dB. It is also visible that too small values of principal components (i.e. $D_{\mathrm{pca}} \leq 6$) impacts the performance at any SNR and L significantly; likewise for too large values ($D_{\mathrm{pca}} \approx 30$) at 0 and 5 dB when L is 10 or 11. This suggests that the new size of the DNN input vector must be chosen after careful analysis as shown in this figure.

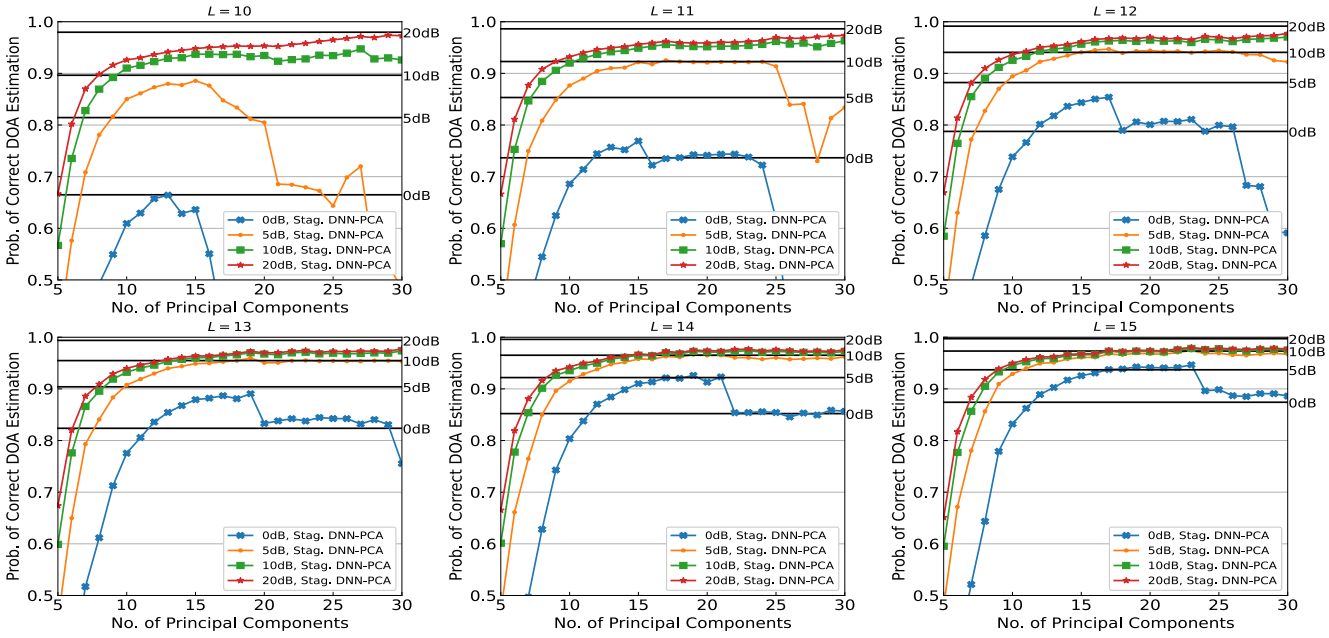


Fig. 7 Comparison of the probability of correct DOA estimation performance of root-MUSIC (solid black lines) and Staggered DNN-PCA for varying number of principal components, or new dimension of input vector after PCA, when the number of antenna elements $L \in [10, 15]$. These methods were tested at 0, 5, 10 and 20 dB.

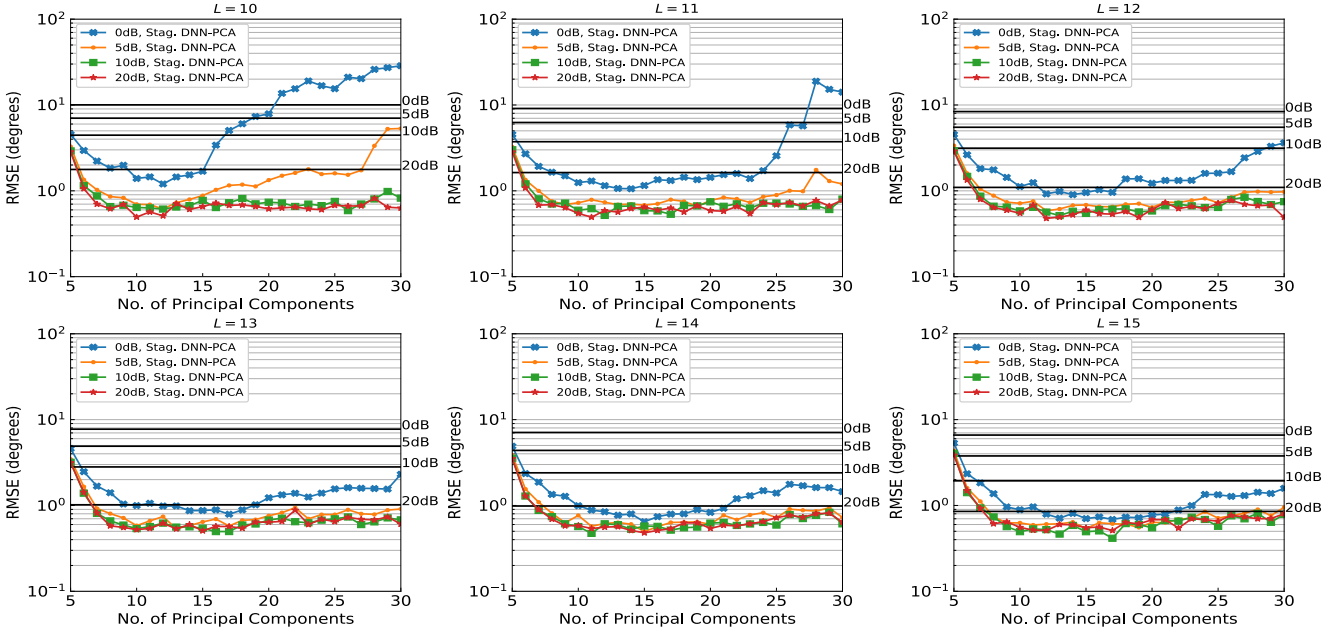


Fig. 8 Comparison of the RMSE performance of root-MUSIC (solid black lines) and Staggered DNN-PCA for varying number of principal components, or new dimension of input vector after PCA, when the number of antenna elements $L \in [10, 15]$. These methods were tested at 0, 5, 10 and 20 dB.

With respect to the RMSE at 0 dB, the optimal number of principal components appears to be slightly different from that corresponding to the probability of correct DOA estimation at each L. This is possibly because the structure of the DNNs (number of hidden layers and units thereof) was optimized in terms of the probability of correct DOA estimation [11], [19], not RMSE. For this reason, we use the optimal value of D_{pca} in terms of the probability of correct DOA estimation (Fig. 7) in the subsequent simulations.

The comparison of the probability of correct DOA estimation and RMSE of Staggered DNN-PCA with those of Staggered DNN and root-MUSIC for varying number of an-

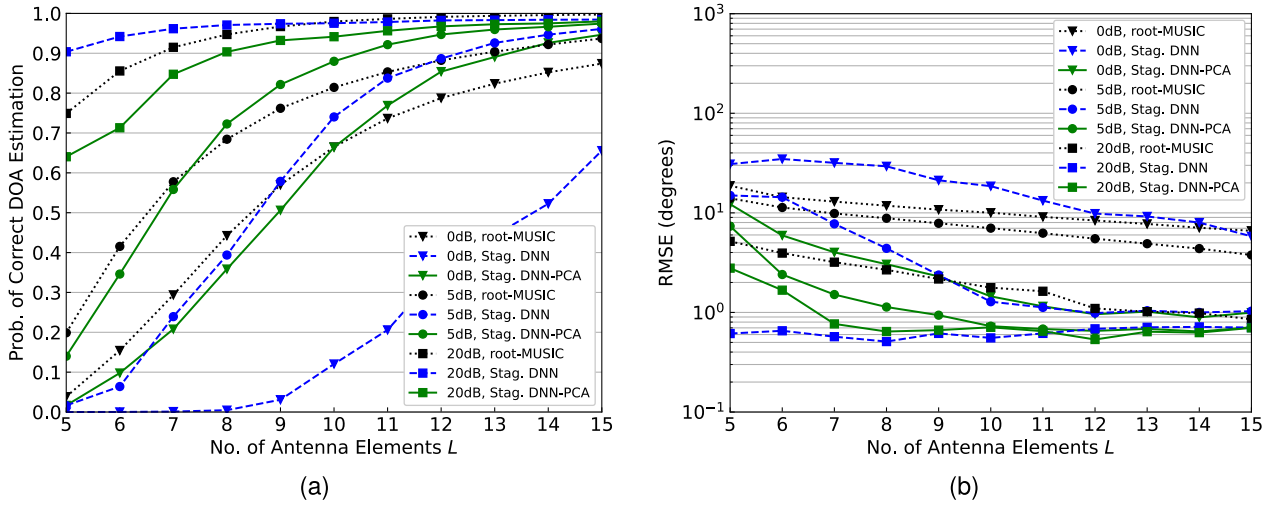


Fig. 9 Performance comparison of 3 DOA techniques while varying the number of antenna elements L: root-MUSIC (dotted black lines), Staggered DNN (dashed blue lines) and Staggered DNN-PCA (solid green lines). These methods were tested at 0, 5 and 20 dB (lower triangle, circle, and square markers, respectively). (a) Probability of correct DOA estimation. (b) RMSE.

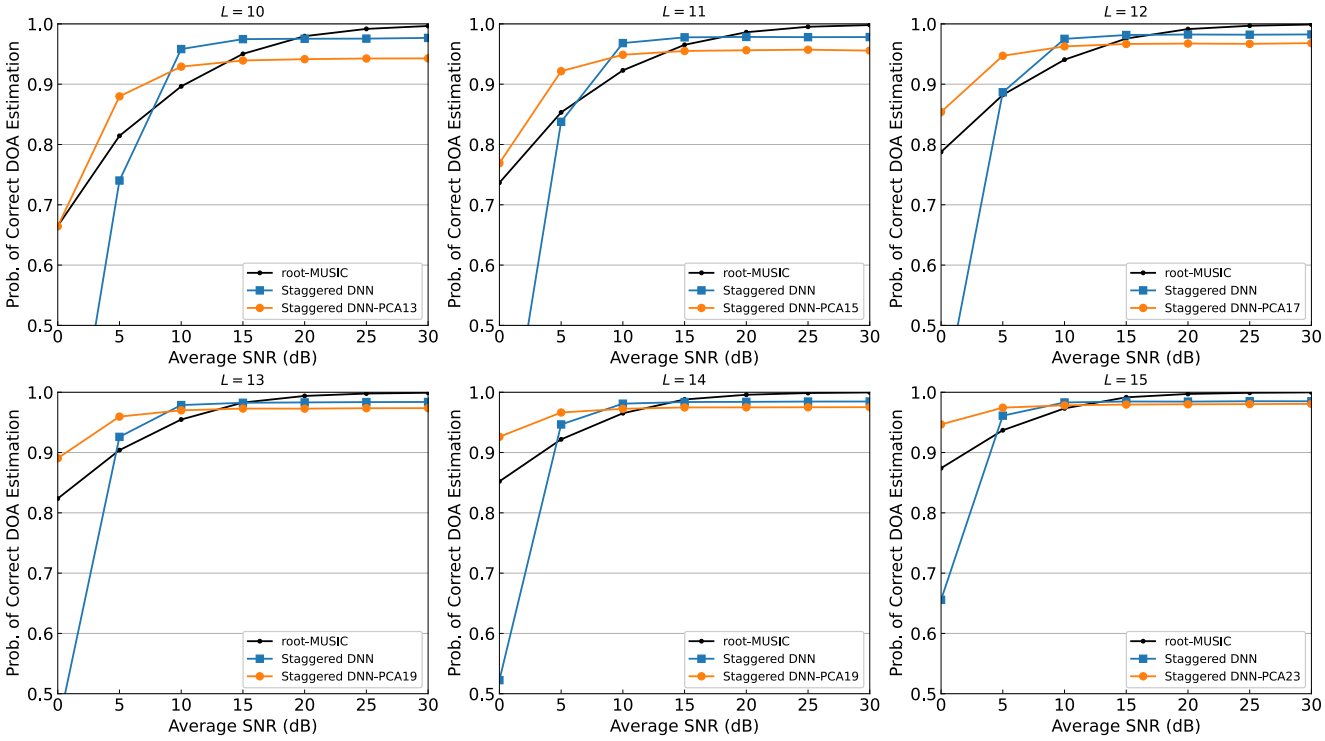


Fig. 10 Comparison of the probability of correct DOA estimation performance of root-MUSIC, Staggered DNN and Staggered DNN-PCAx for varying test SNRs when the number of antenna elements $L \in [10, 15]$. The notation PCAx denotes the number of principal components x chosen.

tenna elements L when the test SNR is 0, 5 and 20 dB is shown in Fig. 9. The number of principal components D_{pca} chosen for each L was the optimal number verified in a graph such as those portrayed in Fig. 7. These values can be found in Table 3, where the dimension reduction $(D_{\text{in}} - D_{\text{pca}})/D_{\text{in}}$ in percentage is also shown. From both figures, the proposed Staggered DNN-PCA is very superior compared to Staggered DNN when the SNR is 0 and 5 dB. Taking as an example the point where $L = 9$, when the input vector dimension has been reduced by approximately 84% (input vector features from 81 to 13), we managed to improve the probability of correct DOA estimation at 0 dB by 12.5 times (raise from 0.04 to 0.5), and the RMSE by −18 dB (reduction from 20 to 2.5 degrees). On the other hand, again we can see that,

Table 3 Dimension reduction $(D_{\text{in}} - D_{\text{pca}})/D_{\text{in}}$ by PCA.

$D_{\text{in}} (= L^2)$	D_{pca}	Dimension Reduction (%)
5^2	8	68.0
6^2	7	80.56
7^2	9	81.63
8^2	11	82.81
9^2	13	83.95
10^2	13	87.0
11^2	15	87.60
12^2	17	88.19
13^2	19	88.76
14^2	19	90.31
15^2	23	89.78

when the SNR is 20 dB, Staggered DNN clearly provides the best performance at any number of antenna elements L. This result indicates once more that PCA does not provide fruitful results at higher SNRs.

Finally, in Fig. 10, we show the probability of correct DOA estimation with respect to the test SNR when L is varied from 10 to 15. When $L \geq 11$, once more it can be seen that not only the proposed Staggered DNN-PCA presents the best performance in all the three methods when the SNR is 0 and 5 dB, but also an input vector dimension reduction of 85% on average (Table 3) is accomplished. In particular the great difference in performance between applying PCA or not should be noted. Even if PCA proves to be ineffective at 10 dB or higher, as the number of antenna elements L increases towards 15, the performance of both Staggered DNNs with and without PCA becomes fairly equal. This could suggest that Staggered DNN-PCA at higher SNRs is more attractive under the condition that L is large. In any case, root-MUSIC still proves to be a stronger algorithm at higher SNRs, given its super-resolution characteristics at high SNR and sufficient large number of snapshots.

5.2 Staggered NRDNN

All parameters for the training of Staggered NRDNN are kept the same as in Tables 1 and 2, except the following:

- Here we only consider $L = 10$;
- The $K = 3$ training and validation DOAs $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$ are generated in a way that they are uniformly distributed within $[\theta_{\min}, \theta_{\max}]$, where (a) $\theta_{\max} - \theta_{\min} = 30°$ and (b) this range is randomly sampled from $[-60.5°, +60.5°]$;
- The number of output layer units of NRDNN-A and NRDNN-B are 31 and 32, respectively.

In Fig. 11, an example of one test spectrum grid when the SNR is 20 dB by using Staggered NRDNN is shown. If the spectrum grid from Staggered DNN alone was used (upper plot in Fig. 11), the DOA detection would be incorrect, where the absolute error of θ_2 would be $|\theta_2 - \hat{\theta}_2| = 0.93°$. As explained in Sect. 4.2, where it is very likely that either DOA

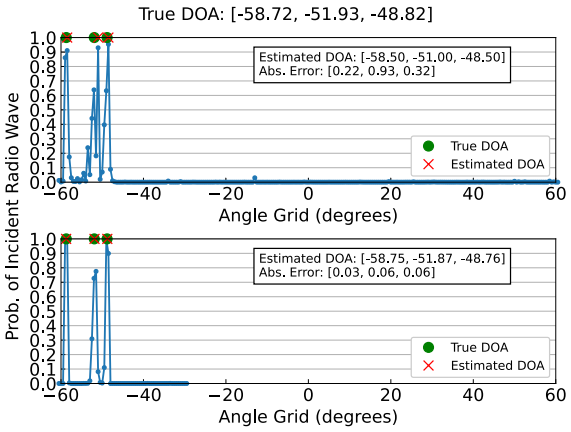


Fig. 11 Example of spectrum grid generation with Staggered DNN (upper figure) and Staggered NRDNN (lower figure) when the SNR is 20 dB and the true DOAs are close within the range of 20°. The green circles and the red X's represent the true DOAs and the estimated DOAs with Neighbors Weighted Average (Sect. 3.3), respectively.

is incorrectly detected when they lie within a range of 20°, and noting that the range of estimated DOAs is less than 20° ($-48.50° - (-58.50°) = 10.00° < 20°$), it can be said that the spectrum grid produced by the appropriate Staggered NRDNN could be more reliable. The selected Staggered NRDNN to produce such spectrum is the one covering the angle bins that fully includes the estimated DOAs, that is, the one that covers the range $[-60°, -30°]$ (first one from the left in Fig. 5(b)). As a result, we obtain the spectrum grid shown in the lower part of Fig. 11. Not only it is a cleaner spectrum, but also it manages to detect all 3 DOAs more precisely, as it can be seen from the considerable drop in the absolute error in the same figure. Therefore, as it will be shown in the next section, this strategy can indeed increase the performance of Staggered DNN at higher SNRs.

5.3 Full System

In Fig. 12, for different test SNRs, the probability of correct DOA estimation and RMSE of the proposed combination of Staggered DNN-PCA and Staggered NRDNN is presented. We compare it again with Staggered DNN and root-MUSIC. The regression-based SNR estimation has been trained and used in the same way as explained in [11]. Here we only consider the case where $L = 10$. The test DOAs were generated as in Table 1.

In terms of probability of correct DOA estimation (Fig. 12(a)), our Staggered NRDNN strategy proposed to cope with close waves shows very good results, especially when the SNR is 20 and 25 dB, where the proposed technique surpasses root-MUSIC performance, while Staggered DNN alone cannot do the same. When the SNR is 30 dB, indeed root-MUSIC still shows better estimation performance; however, our proposed method still manages to perform well. We believe that its use is more attractive than root-MUSIC due to lesser online computational cost, once all necessary DNNs have been offline trained. From the RMSE

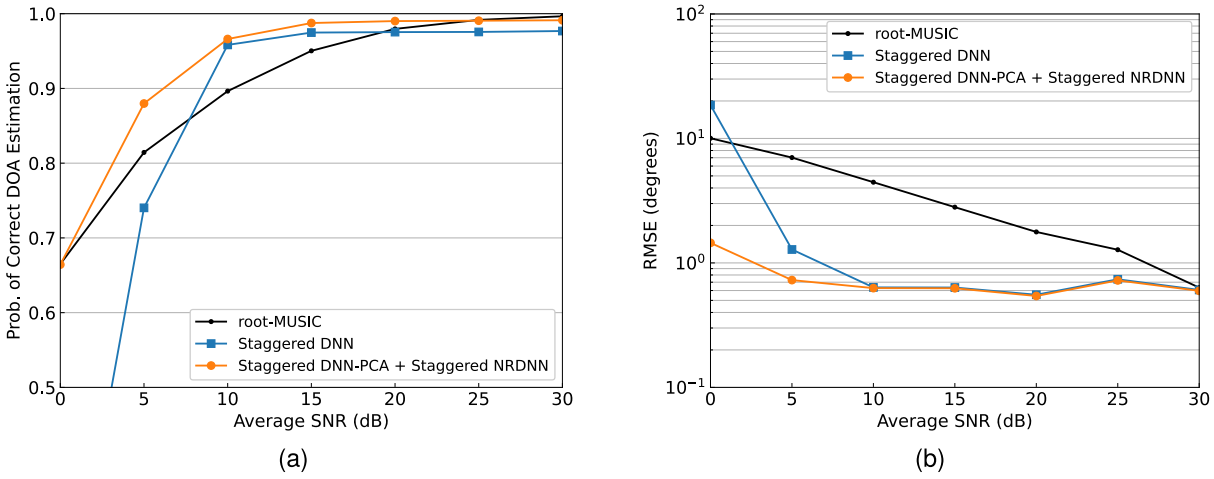


Fig. 12 Performance comparison of root-MUSIC, Staggered DNN and full system (Staggered DNN-PCA + Staggered NRDNN) for varying test SNRs when $L = 10$. The number of principal components chosen for Staggered DNN-PCA was the corresponding optimal value of 13. (a) Probability of correct DOA estimation. (b) RMSE.

in Fig. 12(b), no considerable change is visible when using Staggered NRDNN at 10 dB and over, but this was expected, since the RMSE is an averaging metric and close waves are statistically less common in accordance to our simulation settings.

6. Final Remarks

In this study, we have developed strategies for improving the DOA estimation performance of our previously proposed DNN-based method. Very good results overall surpassing root-MUSIC were achieved in the past when only two radio wave sources were considered. However, as reported in [11], in addition to the fact that accuracy at low SNRs is overwhelmingly poor unless multiple DNNs trained under these conditions are provided, in the event of three radio waves, estimation performance also drops considerably, especially at high SNRs. Consequently, the need to develop schemes that handle these deficiencies was apparent.

Therefore, in this paper we have proposed two separate strategies, each of which tackles these issues at low and high SNRs independently. At low SNRs, we have demonstrated that estimation accuracy is tremendously improved by representing the DNN input vector in a lower dimension (reduction of approximately 85%) by means of PCA, even though this data is generated at a much higher SNR. Additionally, by reducing the size of the input layer, we concurrently manage to reduce the computational cost of DNN. At high SNRs, after noticing that the majority of incorrect estimation cases are due to close waves, we have developed a method where different DNNs specialized in close waves are used instead of the conventional DNN, resulting in a more reliable narrow DOA spectrum grid for subsequent DOA detection. Finally, in order to combine both strategies in a way that such DNN could potentially be deployed in a real scenario, we have used a previously proposed idea [11] of estimating the SNR of the incoming radio waves, so that the appropriate strategy can be switched depending on this SNR. We have obtained great results with this proposed system for the case of three sources with the promise that it could be used instead of root-MUSIC in a bid to acquire better DOA estimation performance while reducing computational cost.

However, further investigation is still necessary before implementation in real scenarios. Despite the brief discussion of the applicability of the proposed methods for higher number of radio wave sources given in the Appendix, more detailed results are still needed. Moreover, study on a less complex SNR estimation module and performance comparison at a) different number of snapshots, b) coherent radio waves, c) uneven power among the incoming waves is necessary.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23H01406.

References

[1] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol.AP-34, no.3, pp.276–280, March 1986. DOI: 10.1109/TAP.1986.1143830.
[2] B.D. Rao and K.V. S. Hari, “Performance analysis of root-MUSIC,” IEEE Trans. Acoust., Speech, Signal Process., vol.37, no.12, pp.1939–1949, Dec. 1989. DOI: 10.1109/29.45540.
[3] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” IEEE Trans. Cogn. Commun. Netw., vol.3, no.4, pp.563–575, Dec. 2017. DOI: 10.1109/TCCN.2017.2758370.
[4] H. Huang, J. Yang, H. Huang, Y. Song, and G. Gui, “Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system,” IEEE Trans. Veh. Technol., vol.67, no.9, pp.8549–8560, Sept. 2018. DOI: 10.1109/TVT.2018.2851783.
[5] M. Chen, Y. Gong, and X. Mao, “Deep neural network for estimation of direction of arrival with antenna array,” IEEE Access, vol.8, pp.140688–140698, Aug. 2020. DOI: 10.1109/

ACCESS.2020.3012582.
[6] D. Hu, Y. Zhang, L. He, and J. Wu, "Low-complexity deep-learning-based DOA estimation for hybrid massive MIMO systems with uniform circular arrays," IEEE Wireless Commun. Lett., vol.9, no.1, pp. 83–86, Jan. 2020. DOI: 10.1109/LWC.2019.2942595.
[7] Z.-M. Liu, C. Zhang, and P.S. Yu, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," IEEE Trans. Antennas Propag., vol.66, no.12, pp.7315–7327, Dec. 2018. DOI: 10.1109/TAP.2018.2874430.
[8] Y. Kase, T. Nishimura, T. Ohgane, Y. Ogawa, D. Kitayama, and Y. Kishiyama, "Fundamental trial on DOA estimation with deep learning," IEICE Trans. Commun., vol.E103-B, no.10, pp.1127–1135, Oct. 2020. DOI: 10.1587/transcom.2019EBP3260.
[9] Y. Kase, T. Nishimura, T. Ohgane, Y. Ogawa, T. Sato, and Y. Kishiyama, "Accuracy improvement in DOA estimation with deep learning," IEICE Trans. Commun., vol.E105-B, no.5, pp.588–599, May 2022. DOI: 10.1587/transcom.2021EBT0001.
[10] D.A. Ando, T. Nishimura, T. Sato, T. Ohgane, Y. Ogawa, and J. Hagiwara, "A proposal of an end-to-end DoA estimation system aided by deep learning," Proc. WPMC 2022, pp.98–103, Oct. 2022. DOI: 10.1109/WPMC55625.2022.10014749.
[11] D.A. Ando, Y. Kase, T. Nishimura, T. Sato, T. Ohgane, Y. Ogawa, and J. Hagiwara, "Deep neural networks based end-to-end DOA estimation system," IEICE Trans. Commun., vol.E106-B, no.12, pp.1350–1362, Dec. 2023. DOI: 10.1587/transcom.2023CEP0006
[12] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzer," Neural Comput 1999, vol.11, no.2, pp.443–482, Feb. 1999. DOI: 10.1162/089976699300016728
[13] J.T.C. Ming, N.M. Noor, O.M. Rijal, R.M. Kassim, and A. Yunus, "Lung disease classification using different deep learning architectures and principal component analysis," Proc. ICBAPS 2018, pp.187–190, July 2018. DOI: 10.1109/ICBAPS.2018.8527385
[14] Z. Chen and L. Pei, "A PCA-BP fast estimation method for broadband two-dimensional DOA of high subsonic flight targets based on the acoustic vector sensor array," Proc. CISP-BMEI 2021, pp.1–6, Oct. 2021. DOI: 10.1109/CISP-BMEI53629.2021.9624358
[15] Y. Liu, H. Chen, and B. Wang, "DOA estimation of underwater acoustic signals based on PCA-kNN algorithm," Proc. CIBDA 2020, pp.486–490, April 2020. DOI: 10.1109/CIBDA50819.2020.00115
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," JMLR vol.12, no.85, pp.2825–2830, 2011.
[17] G.W. Stimson, Introduction to Airborne Radar, 2nd ed., p.183, SciTech Publishing, Mendham, NJ, US, 1998.
[18] S. Ju, Y. Xing, O. Kanhere, and T.S. Rappaport, "Millimeter wave and sub-terahertz spatial statistical channel model for an indoor office building," IEEE J. Sel. Areas Commun., vol.39, no.6, pp.1561–1575, June 2021. DOI: 10.1109/JSAC.2021.3071844.
[19] D.A. Ando, T. Nishimura, T. Sato, T. Ohgane, Y. Ogawa, and J. Hagiwara, "Performance analysis of DNN-PCA for DOA estimation with three radio wave sources," Proc. ISCIT 2023, pp.436–441, Oct. 2023.
[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv:1502.03167v3, March 2015.
[21] D.P. Kingma and J.L. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980v9, Jan. 2017.

Appendix A: Detailed Description of Neighbors Weighted Average

The fully detailed algorithm of "Neighbors Weighted Average" (Sect. 3.3) is presented in Algorithm 1. Its goal is to detect the DOA information from the Staggered DNN output, i.e. the angular spectrum grid estimated by the Staggered DNN, for as many output situations as possible, since spurious bins can hinder proper detection.

Algorithm 1 "Neighbors Weighted Average"

```
1: Parameters: K, t̂, ε ← 0.1, ζ ← 3
2: Result: θ̂
3: while True do
4:     K̂ ← countSources( ε, t̂ );
5:     if K̂ = K then
6:         θ̂ ← NWA( ε, t̂ );
7:         break;
8:     else if K̂ > K then
9:         ε ← ε + 0.1;
10:        if ε > 0.4 then
11:            θ̂ ← peakSearch(K, t̂);
12:            break;
13:        end if
14:    else if K̂ < K then
15:        while True do
16:            θ̂ ← NR-NWA( ε, ζ, t̂ );
17:            if length(θ̂) = K then
18:                break;
19:            else
20:                ζ ← ζ − 1;
21:                if ζ = 0 then
22:                    θ̂ ← peakSearch(K, t̂);
23:                    break;
24:                end if
25:            end if
26:        end while
27:        break;
28:    end if
29: end while
```

The algorithm takes as parameters:

- The number of radio wave sources K, which is considered to be known;
- The output of Staggered DNN $\hat{\mathbf{t}}$ (or estimated spectrum grid);
- The probability threshold ϵ with starting value of 0.1;
- The limit on the number of bins ζ for computation of the weighted average of close DOAs, with starting value of 3.

As previously explained in Sect. 3.3, the ϵ and ζ values are not necessarily optimized; yet, we have achieved great results.

Not always a clean spectrum grid is estimated, with K clearly formed hills. For this reason, the first step (line 4) is to count the number of hills $\hat{K}$ present within the estimated spectrum $\hat{\mathbf{t}}$. The probability threshold ϵ is necessary in this moment. The next step relies on $\hat{K}$:

1. Is $\hat{K} = K$? (lines 5–7)
2. Is $\hat{K} > K$? (lines 8–13)
3. Is $\hat{K} < K$? (lines 14–23)

If case 1, then the weighted average in (9) is simply applied, where all n_{bin} bins within a hill are included in this computation, as shown below:

$$\hat{\theta} = \sum_{i=1}^{n_{\text{bin}}} p_i b_i \Bigg/ \sum_{i=1}^{n_{\text{bin}}} p_i \ . \tag{A·1}$$

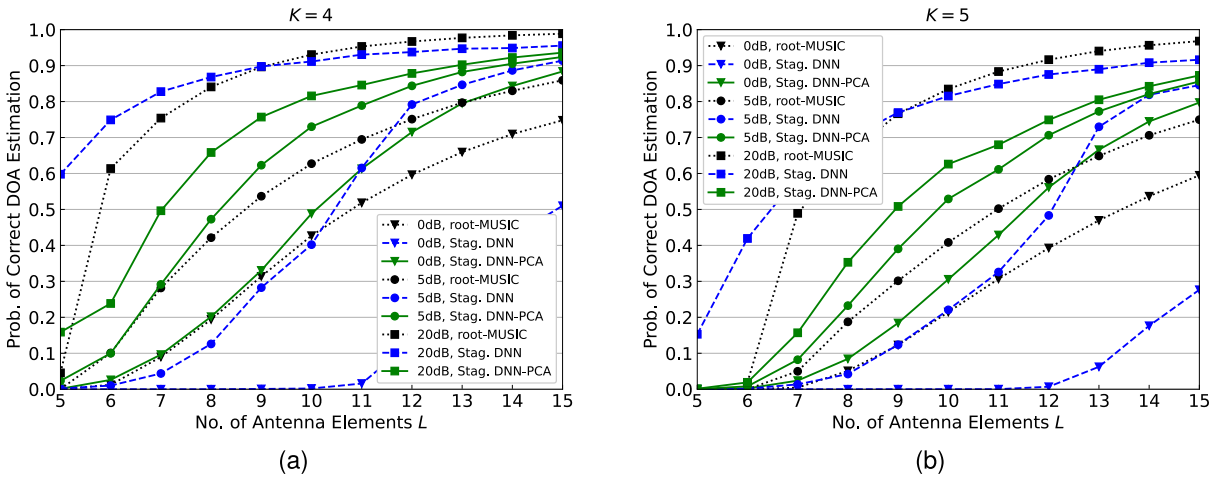


Fig. A· 1 Comparison of the performance in terms of probability of correct DOA estimation of 3 DOA techniques while varying the number of antenna elements L: root-MUSIC (dotted black lines), Staggered DNN (dashed blue lines) and Staggered DNN-PCA (solid green lines). These methods were tested at 0, 5 and 20 dB (lower triangle, circle, and square markers, respectively). (a) $K = 4$. (b) $K = 5$.

Case 2 occurs when there are spurious bins higher than the probability threshold ϵ, resulting in spurious hills that should not be included in the DOA detection. A solution for this problem is to increment the value of ϵ step by step (line 9) until such spurious hills are damped and K clear hills are present for DOA detection. Yet, we set 0.4 as a limit to ϵ. If such a limit is reached and still there are no exact K hills, then we apply a traditional peak search algorithm to the Staggered DNN output $\hat{\mathbf{t}}$ (line 11), where the bins corresponding to the K largest peaks are chosen as the DOA estimates $\hat{\boldsymbol{\theta}}$.

Case 3 most probable occurrence is in the event of close radio waves, which results in overlapping hills. In this case, we need to set a bin limit ζ in order to calculate the weighted average (line 16):

$$\hat{\theta} = \sum_{i=1}^{\zeta} p_i b_i \Big/ \sum_{i=1}^{\zeta} p_i \ . \tag{A· 2}$$

We have verified that $\zeta = 3$ is a good choice. Nevertheless, there are some situations where the number of bins ζ corresponding to one or more DOAs is less than 3. For this reason, we gradually decrement the value of ζ. If this value reaches 0, then again we apply peak search to the Staggered DNN output $\hat{\mathbf{t}}$ (line 22), where the bins corresponding to the K largest peaks are chosen as the DOA estimates $\hat{\boldsymbol{\theta}}$.

Appendix B: Performance of Proposed Methods for Higher Number of Sources

The scope of this study includes primarily the performance analysis of the proposed methods Staggered DNN-PCA in Sect. 4.1 and Staggered NRDNN in Sect. 4.2 for the case of only 3 radio wave sources (i.e. $K = 3$). However, verification of their applicability for higher values of K is also necessary as a fundamental step for future deployment in real-scenario applications. Therefore, in this appendix, we give a brief analysis of the DOA estimation performance of both proposed methods when K is 4 or 5. In practical scenarios with alternating numbers of sources, we envision a dynamic system which consists of several Staggered DNN-PCAs and NRDNNs, each corresponding to each K, that are concurrently deployed according to this K.

B.1 Staggered DNN-PCA

Figure A· 1 shows the extension of the results in Fig. 9(a) when the number of sources K is 4 (Fig. A· 1(a)) and 5 (Fig. A· 1(b)). Here, the optimal number of principal components for each L was found in the same manner as it was done for Fig. 7. All other parameters were kept unchanged. Comparing Fig. A· 1 with Fig. 9(a), the performance of Staggered DNN-PCA appears to be the best at any value of L at 0 and 5 dB; especially when $K = 5$ at 0 dB (see the straight green line with triangle marker in Fig. A· 1(b)), Staggered DNN-PCA excels over root-MUSIC at all L, which contrasts with the case $K = 3$ in Fig. 9(a), where the probability of correct DOA estimation of Staggered DNN-PCA only surpasses that of root-MUSIC when $L \geq 10$.

Figure A· 2 shows the performance of Staggered DNN-PCA at different test SNRs. This is the extension of the results in Fig. 10 for the case $L = 10$. Although the performance of all DOA estimation methods, including root-MUSIC, is degraded as K increases, Staggered DNN-PCA still shows the best probability of correct DOA estimation at 0 and 5 dB. On the other hand, comparing with the performance of Staggered DNN at SNRs of 10 dB or greater, the performance of Staggered DNN-PCA worsens considerably as K increases. We stated in Sect. 5.1 that information on the signal, which is only lightly corrupted by noise at higher SNRs, is lost by applying PCA. Moreover, as K increases, inaccurate DNN outputs are more often produced due to radio waves with close DOA (more to be discussed in the next

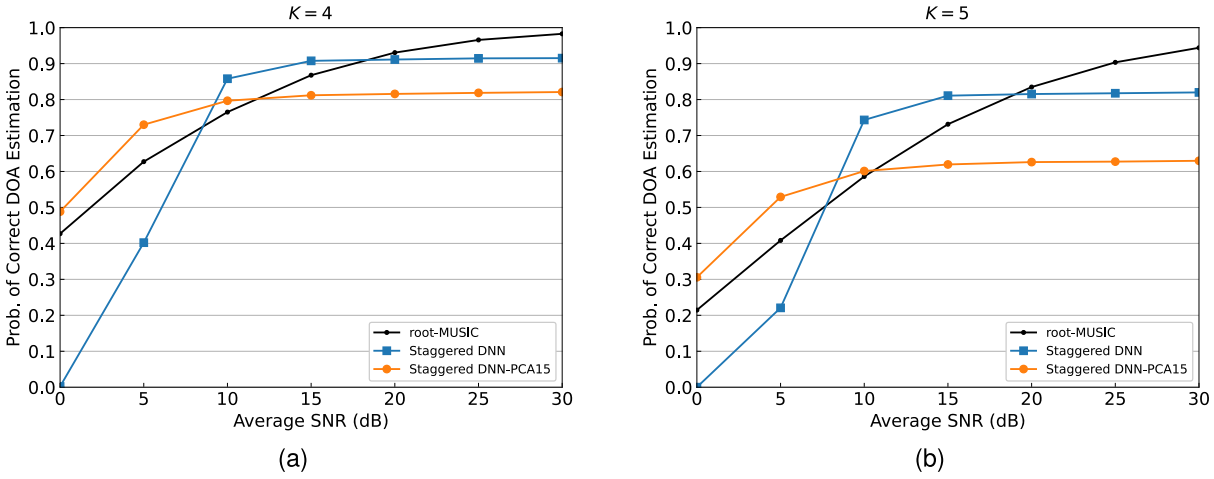


Fig. A·2 Comparison of the probability of correct DOA estimation performance of root-MUSIC, Staggered DNN and Staggered DNN-PCAx for varying test SNRs when the number of antenna elements $L = 10$. The notation PCAx denotes the number of principal components x chosen. (a) $K = 4$. (b) $K = 5$.

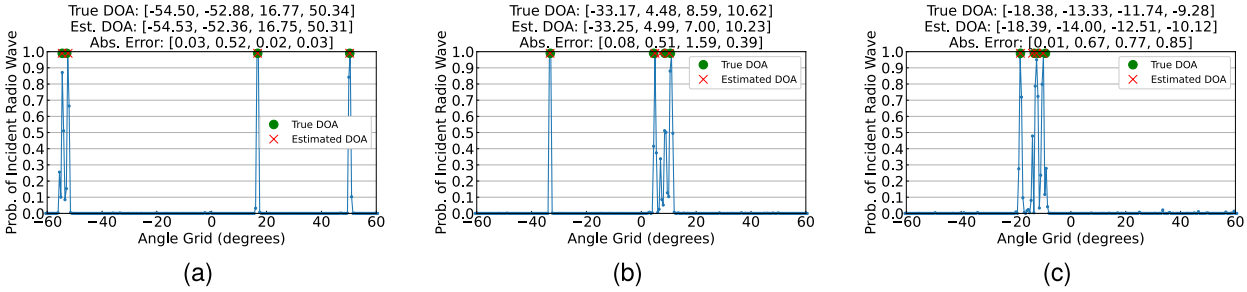


Fig. A·3 Examples of Staggered DNN output when DOA was incorrectly detected at 20 dB for the case $K = 4$. The green circles and the red X's represent the true DOAs and the estimated DOAs, respectively.

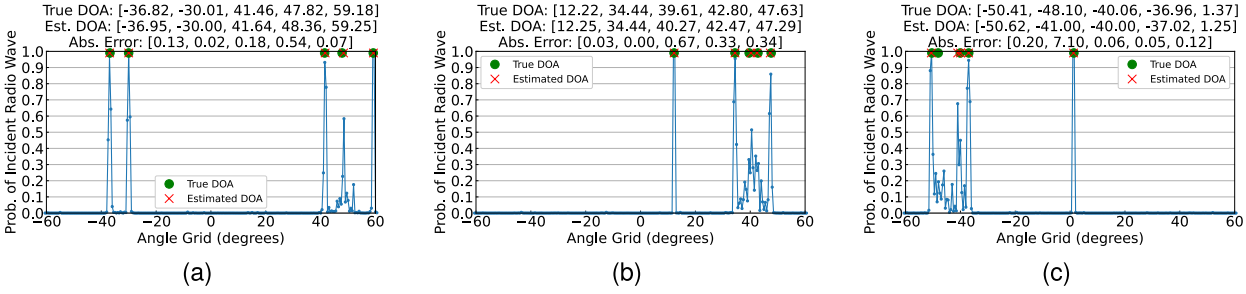


Fig. A·4 Examples of Staggered DNN output when DOA was incorrectly detected at 20 dB for the case $K = 5$. The green circles and the red X's represent the true DOAs and the estimated DOAs, respectively.

section). The combination of both factors are believed to be the reason for significant degradation at higher SNRs as K increases.

In conclusion, applying PCA is still very effective for DOA estimation improvement at lower SNRs when the number of radio wave sources K is 4 and 5.

B.2 Staggered NRDNN

It was explained in Sect. 4.2 and verified in Sect. 5.2 that the proposed Staggered NRDNN is effective in producing more accurate angular spectra; thus, improving DOA estimation performance. However, this approach was designed based on the observation that $K = 3$ radio waves with DOA range of 20° are the main cause of incorrect DOA estimation at higher SNRs with Staggered DNN. Observing the results presented in Figs. A·3 and A·4, which show examples of the output (i.e. angular spectrum) of Staggered DNN when DOA estimation was incorrect at 20 dB for the case $K = 4$ and $K = 5$, respectively, we concluded that, as K increases,

the possible patterns of angular spectrum corresponding to incorrect DOA also increases. For instance, as opposed to the case $K = 3$, we can observe from Figs. A· 3 and A· 4 such patterns of angular spectrum where not only all K peaks, but also $K - K'$ peaks are in close range. Here, K' is the number of radio waves apart from the close range of 20°. In fact, we verified that, by applying Staggered NRDNN as described in Sect. 4.2 to the cases $K = 4$ and $K = 5$, no significant improvement in DOA performance was achieved. However this was an expected result, since this method was designed for $K = 3$. We believe that the redesign and retrain of this method based on different K can result in more accurate DOA estimation. Since this investigation is out of the scope of this paper, we leave it as future work.

Daniel Akira Ando received the B.E. degree in communication networks engineering from University of Brasilia, Brazil, in 2018 and the M.E. degree in media networks engineering from Hokkaido University, Japan, in 2021. He is currently pursuing the Ph.D. degree at Hokkaido University, Japan. His research interests are in MIMO signal processing for wireless communications. He received the IEICE RCS Young Researcher Award in 2020.

Toshihiko Nishimura received the B.S. and M.S. degrees in physics and Ph.D. degree in electronics engineering from Hokkaido University, Sapporo, Japan, in 1992, 1994, and 1997, respectively. Since 1998, he has been with Hokkaido University, where he is currently a Professor. His current research interests are in MIMO systems using smart antenna techniques. He received the Young Researchers' Award of IEICE in 2000, the Best Paper Award from IEICE in 2007, and TELECOM System Technology Award from the Telecommunications Advancement Foundation of Japan in 2008, the best magazine paper award from IEICE Communications Society in 2011, and the Best Tutorial Paper Award from the IEICE Communications Society in 2018. He is a member of the IEEE.

Takanori Sato was born in Hokkaido, Japan, in 1992. He received his Ph.D. degree in the field of media and network technologies from Hokkaido University, Japan, in 2018. He was a Research Fellow of Japan Society for the Promotion of Science (JSPS) from 2017 to 2019. In 2019, he moved to University of Hyogo as an assistant professor. He is currently an associate professor in Hokkaido University. His research interests include the theoretical and numerical studies of optical fibers and photonic circuits using the coupled mode theory and the finite element method. He is a member of the Japan Society of Applied Physics (JSAP), Institute of Electrical and Electronics Engineers (IEEE), and the Optical Society of America (OSA).

Takeo Ohgane received the B.E., M.E., and Ph.D. degrees in electronics engineering from Hokkaido University, Sapporo, Japan, in 1984, 1986, and 1994, respectively. From 1986 to 1992, he was with Communications Research Laboratory, Ministry of Posts and Telecommunications. From 1992 to 1995, he was on assignment at ATR Optical and Radio Communications Research Laboratory. Since 1995, he has been with Hokkaido University, where he is currently a Professor. During 2005–2006, he was at Centre for Communications Research, University of Bristol, U.K., as a Visiting Fellow. His research interests are in MIMO signal processing for wireless communications. He received the IEEE AP-S Tokyo Chapter Young Engineer Award in 1993, the Young Researchers' Award of IEICE in 1990, the Best Paper Award from IEICE in 2007, TELECOM System Technology Award from the Telecommunications Advancement Foundation of Japan in 2008, the Best Magazine Paper Award from IEICE Communications Society in 2011, and the Best Tutorial Paper Award from IEICE Communications Society in 2018. He is a member of the IEEE.

Yasutaka Ogawa received the B.E., M.E., and Ph.D. degrees from Hokkaido University, Sapporo, Japan, in 1973, 1975, and 1978, respectively. Since 1979, he has been with Hokkaido University, where he is currently a Professor Emeritus. During 1992–1993, he was with ElectroScience Laboratory, the Ohio State University, as a Visiting Scholar, on leave from Hokkaido University. His professional expertise encompasses super-resolution estimation techniques, applications of adaptive antennas for mobile communication, multiple-input multiple-output (MIMO) techniques, and measurement techniques. He proposed a basic and important technique for time-domain super-resolution estimation for electromagnetic wave measurement such as antenna gain measurement, scattering/diffraction measurement, and radar imaging. Also, his expertise and commitment to advancing the development of adaptive antennas contributed to the realization of space division multiple accesses (SDMA) in the Personal Handy-phone System (PHS). He received the Yasujiro Niwa Outstanding Paper Award in 1978, the Young Researchers' Award of IEICE in 1982, the Best Paper Award from IEICE in 2007, TELECOM system technology award from the Telecommunications Advancement Foundation of Japan in 2008, the Best Magazine Paper Award from IEICE Communications Society in 2011, the Achievement Award from IEICE in 2014, and the Best Tutorial Paper Award from IEICE Communications Society in 2018. He also received the Hokkaido University Commendation for excellent teaching in 2012. He is a Life Fellow of the IEEE.

Junichiro Hagiwara received the B.E., M.E., and Ph.D. degrees from Hokkaido University, Sapporo, Japan, in 1990, 1992, and 2016, respectively. He joined the Nippon Telegraph and Telephone Corporation in April 1992 and transferred to NTT Mobile Communications Network, Inc. (currently NTT DOCOMO, INC.) in July 1992. Later, he became involved in the research and development of mobile communication systems. His current research interests are in the application of stochastic theory to the communication domain. He was a visiting professor at Hokkaido University from 2018 to 2023. He is a member of the IEEE.

PAPER

Effects of Site Diversity Techniques on the Rain Attenuation in Ku-Band Satellite Communications Links According to the Kind of Rain Fronts

Yasuyuki MAEKAWA[†a)], **Yoshiaki SHIBAGAKI**[†], ***Members***, ***and*** **Tomoyuki TAKAMI**[††], ***Nonmember***

SUMMARY The effects of site diversity techniques on Ku-band rain attenuation are investigated using two kinds of simultaneous BS (Broadcasting Satellite) signal observations: one was conducted among Osaka Electro-Communication University (OECU) in Neyagawa, Kyoto University in Uji, and Shigaraki MU Observatory in Koka for the past ten years, and the other was conducted among the headquarter of OECU in Neyagawa and their other premises in Shijonawate and Moriguchi for the past seven years, respectively. The site diversity effects among these sites with horizontal separations of 3–50 km are found to be largely affected by the passage direction of rain areas characterized by each rain type, such as warm, cold, and stationary fronts or typhoon and shower. The performance of the site diversity primarily depends on the effective distance between the sites projected to the rain area motions. The unavailable time percentages are theoretically shown to be reduced down to about 61–73% of the ITU-R predictions by choosing a pair of the sites aligned closest to the rain area motion in the distance of 3–50 km. Then, we propose three kinds of novel site diversity methods that choose the pair of sites based on such as rain type, rain front motion, or rain area motion at each rainfall event, respectively. As a result, the first method, which statistically accumulates the average passage directions of each rain type from long-term observations, is even useful for practical operations of the site diversity, because unavailable time percentages are reduced down to about 75–85% compared with the theoretical limit of about 61–73%. Also, the third method based on the rain area motion directly obtained from the three-site observations yields the reduction in unavailable time percentages close to this theoretical limit.

key words: *satellite communications, rain attenuation, Ku band, rain area motion, wind velocity, rain fronts*

1. Introduction

Site diversity techniques are frequently used to mitigate rain attenuation effects that are significant in satellite communications using frequency of higher than 10 GHz. So far, a number of rain attenuation measurements have been conducted between two sites, to investigate the site diversity effects. Also, the prediction methods for cumulative time percentages of the site diversity effects using the distance between the two sites are well established in terms of the improvement of their joint time percentages [1], [2].

In our previous study [3], the effects of rain area motions on the site diversity performance were discussed among the three sites in the area of 20–50 km. These measurements were conducted at Osaka Electro-Communication University (OECU) in Neyagawa, Osaka, Research Institute of Sustainable Humanosphere (RISH) in Uji, Kyoto, and Shigaraki MU Observatory of Kyoto University (MU) in Koka, Shiga for the past ten years from Sept. 2002 to July 2011. Then, the site diversity performance is shown to be improved, when a pair of the two sites is chosen to be aligned closest to the rain area motion.

In this study, the site diversity effects are further investigated in relation to the rain area motions of various rain types, between OECU and other two sites in the narrower area of 3–8 km, to see their horizontal structures in more detailed scale. In this experiment the Ku-band satellite signal attenuation was measured for the past seven years from July 2005 to July 2011, at two additional sites located at the other premises of OECU in Moriguchi (Mori) and Shijonawate (Shijo), Osaka, which are both a few kilometers away from the headquarter of OECU in Neyagawa (Neya) [4]. The effects of the rain area motions are evaluated in a wide range of the distances from 3 to 50 km in terms of the improvement of the site diversity performance. Specifically, the effects of novel site diversity methods are evaluated by choosing the pair of the two sites based on such as rain type, rain front motion, or rain area motion, respectively.

Then, the reduction in unavailable time percentages compared to the ITU-R predictions is quantitatively calculated, using geometric figures of the three observational sites, and the time percentages are shown to be reduced down to about 61–73% compared to the conventional ITU-R predictions. It is also shown that even with two stations, which is easier to operate compared to the site diversity that uses all three stations, we can obtain the reduction in unavailable time percentages considerably similar to the case with three stations, when the appropriate two stations are chosen.

Furthermore, it is known that the site diversity effects depend on the arrangement between two stations and the arrival direction of the radio waves [5]. We point out that this problem also occurs between three stations in the narrow area with a distance of 10 km or less and affects the observational results.

In this paper, to discuss the site diversity effects, the yearly time percentages observed at a single site are denoted by P_1, while the yearly joint time percentages given by

Manuscript received February 12, 2024.
Manuscript revised May 12, 2024.
Manuscript publicized July 18, 2024.
[†]Osaka Electro-Communication University, Neyagawa-shi, 572-8530 Japan.
[††]Osaka Electro-Communication University, Shijonawate-shi, 575-0063 Japan.
a) E-mail: maekawa@osakac.ac.jp
DOI: 10.23919/transcom.2024EBP3030

the site diversity between two stations are expressed by P_2. Moreover, the yearly joint time percentages of the site diversity between the two stations aligned closest to the rain area motion is represented by P'_2. Then, we define the "reduction rate of unavailable time percentages" as P'_2/P_2, when the appropriate two stations are chosen for the site diversity considering the rain area motion. Also in this study, the samples of rain attenuation observed on rainy days associated with each rain type are referred to as "rainfall events".

2. Observation Methods

At the three sites of OECU, RISH and MU, the Ku-band broadcasting satellite (BS) signals (11.84 GHz, circular polarization, elevation angle 41.3°) were continuously observed from 2002 to 2012. At RISH in Uji, however, the Ku-band signal (12.74 GHz, horizontal polarization, elevation angle 48.5°) of Superbird C was observed up to July 2005 [3]. On the other hand, at the two nearby premises of OECU in Mori and Shijo, BS signals were also continuously measured from 2005 to 2012. These signal levels are recorded every second by personal computers equipped with 16 bit AD converters, and averaged over 1 min for further analyses. Also, rainfall rate is recorded at 1 min interval with the resolution of 0.1 mm in all of these sites.

In the wide area of 20–50 km, RISH in Uji, Kyoto is located 23.3 km northeast (16.0 km, 16.9 km) from OECU in Neyagawa, Osaka, while MU in Koka, Shiga is located 45.9 km east northeast (44.2 km, 12.4 km) from OECU. In the narrow area of 3–8 km, on the other hand, the sites at Mori and Shijo premises are located 5.6 km southwest (−5.0 km, −2.6 km) and 3.9 km southeast (3.2 km, −2.1 km) from OECU in Neyagawa, respectively. These locations of two kinds of three-site BS signal observations in both wide and narrow areas are illustrated in Fig. 1 [6].

3. Example of Observations

Figure 2 shows an example of rain attenuation observed at the three sites of OECU, MU, and RISH in the wide area, on July 10, 2007. In Fig. 2(a), we can see that the attenuation of 5 dB (dashed line) or more occurred during 5:30–6:40 LT at each site in the order of OECU (dark blue line), RISH (green line), and MU (red line). The rainfall rate of about 14 mm/h was recorded at OECU during 5:30–5:50 LT.

In Fig. 2(b), the site diversity effects are calculated for each combination of the three sites, which is switched between OECU and MU (red line), OECU and RISH (dark blue line), and RISH and MU (green line), respectively. Note that after the site diversity is performed between these two stations, the attenuation exceeding 5 dB (dashed line) is never observed.

In Fig. 2(c), on the other hand, cross-correlation functions of the rain attenuation are calculated between OECU and the other sites. The red and dark blue lines indicate the results obtained from the combination of OECU and MU, and that of OECU and RISH, respectively. The lag times

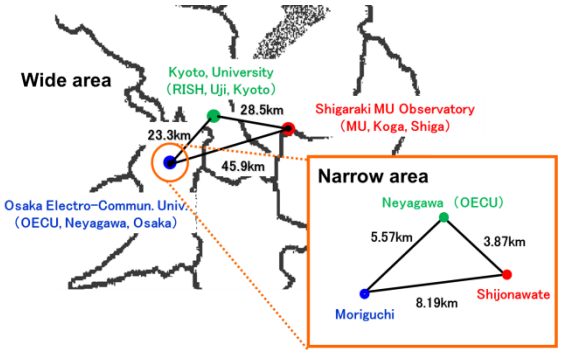


Fig. 1 Locations of wide and narrow area three-site observations of the BS signals [6].

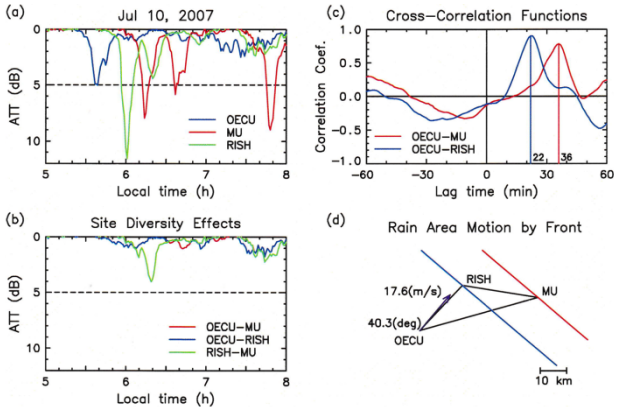


Fig. 2 Example of rain attenuation and rain area motion estimated in OECU, MU, and RISH.

obtained from these peaks indicate that the attenuation occurred 22 min and 36 min later at RISH and MU, respectively, than at OECU. Also, note that the cross-correlation coefficients at these peaks are as high as 0.8, even though the distances among the three sites are from 20 to 50 km.

Figure 2(d) shows velocity (arrow) of the rain area estimated from the time differences in attenuation occurrence among the three sites, as well as their geographical relationship with OECU [3]. Dark blue and red lines indicate the positions of the rain front inferred at RISH and MU, respectively, which passed over them in this order. The rain area associated with the front is shown to move northeastward at a speed of 17.6 m/s. The direction of the motion is found to be 40.3°. For the past ten years from 2002 to 2011, 378 rainfall events indicating samples of such rain area motions were similarly obtained from the sites in OECU, MU, and RISH. Also, the direction of rain area motion is hereinafter indicated by clockwise from the north.

Figure 3 also shows an example of rain attenuation observed at the three sites of OECU (Neya), Mori, and Shijo in the narrow area, on July 10, 2007. In Fig. 3(a), the attenuation of nearly 5 dB (dashed line) was similarly found at these three sites around 5:30–5:50 LT. Also, Fig. 3(b) presents the site diversity effects between Mori and Shijo (red line), Mori and Neya (dark blue line), and Neya and Shijo (green line),

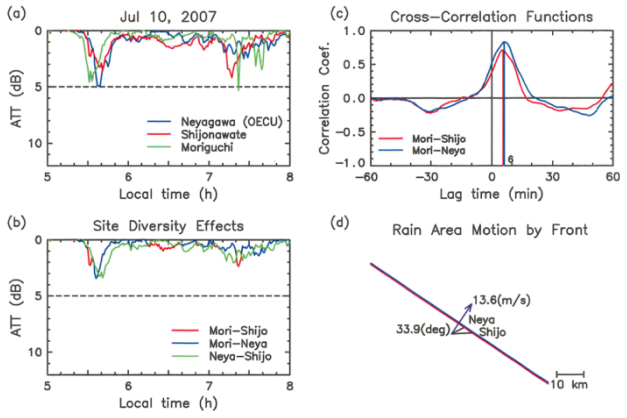


Fig. 3 Example of rain attenuation and rain area motion estimated in Neyagawa (OECU), Shijonawate, and Moriguchi.

Table 1 Speed and direction of the rain area motion on July 10, 2007.

Measurement method	Speed	Direction
Wide area three-site observation	17.6 m/s	40.3°
Narrow area three-site observation	13.6 m/s	33.9°
Weather charts	12.1 m/s	47.0°

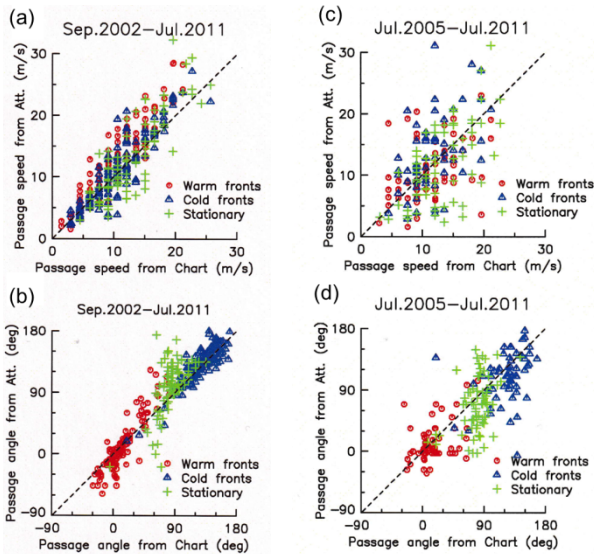


Fig. 4 Scatter plots of speed and direction of the rain area motion between the three-site observations and weather charts in the wide area (left, a, b) and the narrow area (right, c, d), respectively.

respectively, among the three sites. After the site diversity, the attenuation exceeding 5 dB (dashed line) is not observed in the narrow area either.

The cross-correlation functions in Fig. 3(c) indicate that lag times of 6 min exist between Mori and Shijo as well as Mori and Neya, and that the cross-correlation coefficients are nearly 0.8. Thus, Fig. 3(d) reveals that the rain area associated with the front moves northeastward at a speed of 13.6 m/s, with the direction of 33.9°.

On July 10, 2007, on the other hand, the weather charts published by Japan Meteorological Agency indicate that a warm front similarly passed northeastward over the Kansai area including Osaka, Kyoto, and Shiga. The speed and direction of the rain area described in terms of the velocity perpendicular to the warm front of the weather charts are 12.1 m/s and 47°, respectively. These values are estimated from the weather charts cited every 12 h in newspapers [3], and in fairly good agreement with the wide and narrow area observations obtained from Fig. 2(d) and Fig. 3(d), respectively. The speed and direction of the rain area obtained from the wide and narrow areas and the weather charts are summarized in Table 1.

For the past seven years from 2005 to 2011, total 186 rainfall events indicating such rain area motions were similarly obtained from the nearby sites in OECU, Shijo, and Mori. These results also indicate fairly good agreement with those obtained from OECU, MU, and RISH, together with the weather charts as will be shown in the next chapter.

4. Speed and Direction of Each Rain Type

Figure 4 shows scatter plots between the wide area three-site observations and the weather charts in the left side of the diagrams, for (a) passage speeds and (b) directions of warm fronts (red), cold fronts (dark blue), and stationary fronts (green), respectively. In these plots, 344 rainfall events were obtained at OECU, MU, and RISH from Sep. 2002 and July 2011 for warm, cold, stationary fronts, other than typhoons and showers that are not accompanied by rain fronts. Figure 4 similarly depicts scatter plots between the narrow area three-site observations and weather charts in the right side, for (c) passage speeds and (d) directions of each rain front. In the narrow area observations, 159 rainfall events were obtained at Neya (OECU), Shijo, and Mori from July 2005 and JulY 2011 except for typhoons and showers.

The speed and direction of warm and cold fronts are similarly obtained from those perpendicular to the front lines as illustrated in Figs. 2(d) and 3(d). The motion of rain areas for stationary fronts is rather inferred from small or medium-size extratropical cyclones (low pressures) that move along the front lines in the weather charts published every 12 h in newspapers [3].

It is seen from Fig. 4 that the speeds and directions of the rain areas estimated from the three-site observations agree fairly well with those of rain fronts or extratropical cyclones directly detected on the weather charts. The correlation coefficients between the three-site observations and weather charts are nearly 0.9 in the wide area, while they are decreased down to nearly 0.6 in the narrow area. The degradation of cross-correlation is possibly due to short time difference in the peaks of attenuation by only a few minutes between them, as shown in Fig. 3(c). Also, the effects of local wind velocities to the narrow area observations should be considered in Osaka plain along Yodogawa (Yodo river), such as land and see breeze [7]. Thus, the three-site observations, as a whole, well represent the motion of rain areas associated with warm and cold fronts or extratropical cyclones in the case of stationary fronts.

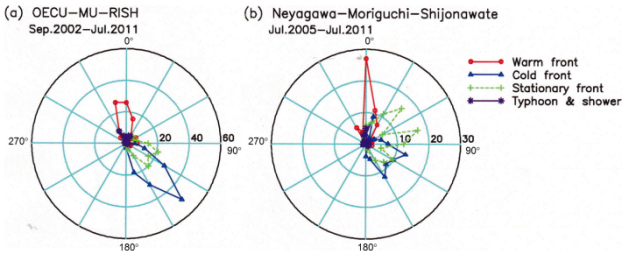


Fig. 5 Distributions of the directions of the rain area motions estimated in the (a) wide and (b) narrow area three-site observations for each rain type over the entire observation period.

Table 2 Average direction of rain area motions in the wide area.

Rain type	Average direction	Event No.
Total	76.86 °	378
Warm front	13.47 °	114
Cold front	132.49 °	145
Stationary front	97.55 °	85
Typhoon and Shower	0.46 °	34

Table 3 Average direction of rain area motions in the narrow area.

Rain type	Average direction	Event No.
Total	64.41 °	186
Warm front	16.06 °	32
Cold front	108.08 °	46
Stationary front	71.28 °	81
Typhoon and Shower	26.68 °	27

Next, Fig. 5 shows distributions of the directions of rain area motions estimated by the two kinds of three-site observations for each rain type including typhoons and showers over the entire observation period. Fig. 5(a) shows the results of 378 rainfall events obtained from OECU, MU, and RISH between Sept. 2002 and July 2011 at 30° intervals. Similarly Fig. 5(b) shows the results of 186 rainfall events obtained from Neya (OECU), Mori, and Shijo between July 2005 and July 2011.

Although there is a slight bias in the angular distributions between the two three-site observations, the rain area, in general, moves from south to north in the warm fronts, whereas it moves, on an average, from northwest to southeast in the cold fronts, and from west to east in the stationary fronts, respectively. In addition, the direction of movement of these rain areas coincides well with the direction perpendicular to the warm and cold fronts, while it coincides with the direction of the low pressure along the stationary fronts, as was shown in Fig. 4. The rain area of typhoons and shower, as a whole, moves from south to north, though the number of rainfall events is small. Tables 2 and 3 summarize the average directions of rain area motions in the rainfall events for total and each rain type, in the wide and narrow area three-site observations, respectively.

5. Site Diversity Effects for Each Rain Type

Figure 6 depicts the cumulative time percentages of the rain attenuation obtained at OECU (dark blue), MU (red), and RISH (green) in the wide area of 20–50 km from Sept. 2002

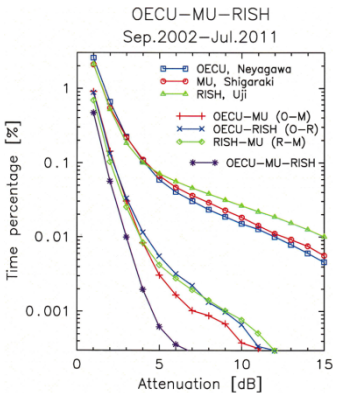


Fig. 6 Cumulative time percentages at OECU, MU, and RISH, and joint time percentages between these sites.

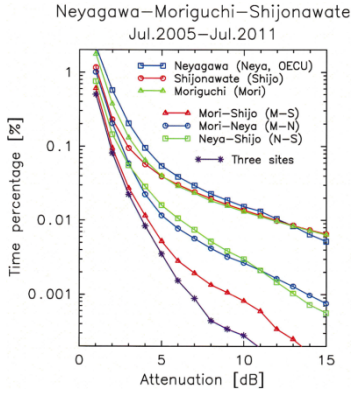


Fig. 7 Cumulative time percentages at Neyagawa, Shijonawate, and Moriguchi, and joint time percentages between these sites.

to July 2011. Also, Fig. 6 shows the results of site diversity effects numerically calculated among the three sites in the lower part of the diagram. The site diversity effects are here evaluated for each combination of the three sites, which is switched between OECU and MU (red, O-M), OECU and RISH (dark blue, O-R), and RISH and MU (green, R-M), respectively, as well as among the three sites (purple).

Figure 7 similarly depicts the cumulative time percentages of the rain attenuation obtained at Neyagawa (Neya, dark blue), Shijonawate (Shijo, red), and Moriguchi (Mori, green) in the narrow area of 3–8 km from July 2005 to July 2011, Also, Fig. 7 shows the results of the site diversity effects numerically calculated among the three sites, respectively. The results are here presented between Mori and Shijo (red, M-S), Mori and Neya (dark blue, M-N), and Neya and Shijo (green, N-S), respectively, as well as among the three sites (purple).

It can be seen from Fig. 6 that between the two sites in the wide area with a distance of 20 km or more, the time percentages of rain attenuation of more than 4 dB are decreased by one order or more due to the site diversity effects. On the other hand, Fig. 7 shows that between the two sites in the narrow area with a distance of 10 km or less, site diversity effects exceeding one order of time percentages are not

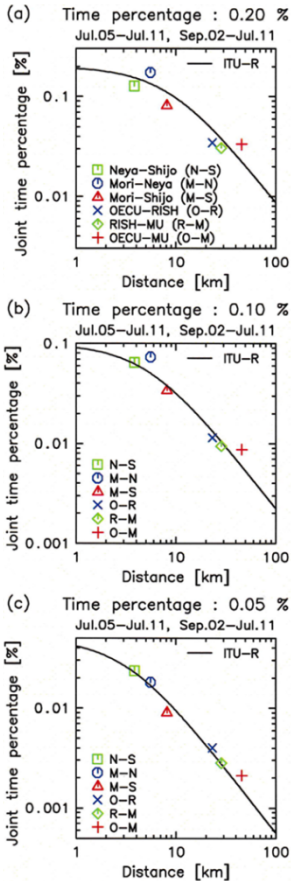


Fig. 8 Joint time percentages for all pairs of the sites in the distance of 3–50 km. The results are shown for the time percentages for (a) 0.2, (b) 0.1, and (c) 0,05% of the single site, respectively [6].

found even for the attenuation of about 12 dB or more, except for Moriguchi and Shijo with the longest distance (red, M-S).

These features of the site diversity effects are then depicted in terms of distance dependence [6]. In Fig. 8, the joint time percentages P_2 obtained for all pairs of the sites are plotted in the order from short to long geographical distances, i.e., N-S (green), M-N (dark blue), and M-S (red) in the narrow area, and O-R (dark blue), R-M (green), and O-M (red) in the wide area, respectively.

The results are presented for the original single-site time percentages P_1 of (a) 0.2, (b) 0.1, and (c) 0.05%, which correspond to the attenuation of 2.2, 3.0, and 4.4 dB for the narrow area in Fig. 7, and the attenuation of 2.9, 4.0 and 5.5 dB for the wide area in Fig. 6, respectively. The attenuation of each time percentage is selected from the lowest value among the three single sites in each area. Specifically, 2.2, 3.0, and 4.4 dB are all taken from Shijonawate station in Fig. 7. On the other hand 2.9 and 4.0 dB are taken from RISH station, while 5.5 dB is taken from OECU station in Fig. 6.

Thin lines indicate joint time percentages predicted by the ITU-R recommendations [2] for the corresponding time percentages of 0.2, 0.1, and 0.05%, respectively. According to the ITU-R methods, the relationship between the joint time percentages P_2 after the site diversity and the original time percentages P_1 before the site diversity is expressed by the improvement factor I as follows

$$I = \frac{P_1}{P_2} = \frac{1}{(1+\beta^2)}\left(1 + \frac{100\beta^2}{P_1}\right) \cong 1 + \frac{100\beta^2}{P_1} \quad (1)$$

$$\beta^2 = 10^{-4} d^{1.33} \quad (2)$$

Thus, as was shown in our previous study [3], the joint time percentages are decreased as the distance between the sites are increased according to the ITU-R recommendations. The time percentages of (a) 0.2% and (b) 0.1%, however, tend to indicate slightly higher joint time percentages at M-N and O-M. In the case of M-N (Mori-Neya), the satellite azimuth angle (220.1°) approaches the alignment of these sites in the distance of 5.57 km which is comparable to the equivalent path length of rain attenuation. So, the site diversity effects seem to be reduced as predicted by the ITU-R recommendations [2].

The conventional study on the baseline between the two stations and the direction of radio wave from a satellite using the radar data of the rainfall intensity has reported that there are both cases that the site diversity effects are reduced and not reduced when the baseline coincides with the satellite azimuth angle [5]. In the present observation, however, the distance of the stations M-N is 5.57 km, which is considerably shorter than 10 km of the two stations in the conventional study, so the effect of the satellite azimuth angle seems to be more prominent.

In the case of OM (OECU-MU), on the other hand, the alignment of these sites tends to become perpendicular to the passage direction of the cold fronts (132.49°) as shown in Table 2. This alignment seems to reduce effective distance along the passage direction, yielding the higher joint time percentages in spite of the longer distance of 45.9 km [3].

Next, the site diversity effects are examined for the four rain types classified in Tables 2 and 3. Their joint time percentages are then similarly calculated for the attenuation of 2.9, 4.0 and 5.5 dB in the wide area, which are equivalent to the yearly time percentages of 0.2, 0.1 and 0.05% for the single site, respectively. The attenuation of each time percentage is selected from the lowest value among the three sites for the single site attenuation similarly to Fig. 8. Figure 9 indicates the distances of all pairs of the three sites (O-R, R-M, O-M) in the left side, and the joint time percentages of the site diversity effects in the right side for each rain type of (a) warm, (b) cold, and (c) stationary fronts and (d) typhoon and shower, respectively. As for (a) the warm front, however, the site diversity effects of the time percentage of 0.05% are not obtained for the pair of O-R, because the number of data is very sparse.

As concerns the distances of the two sites depicted in the left-side plots, their average lengths projected to the passage direction of rain area motion, hereinafter referred to as "effective distance", are indicated (red), together with their

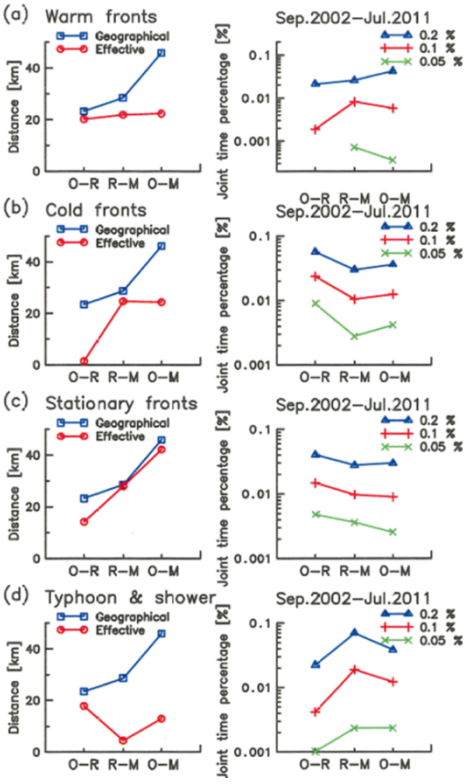


Fig. 9 Site diversity effects of each rain type for the cumulative time percentages of 0.2–0.05% obtained in the wide area of 20–50 km.

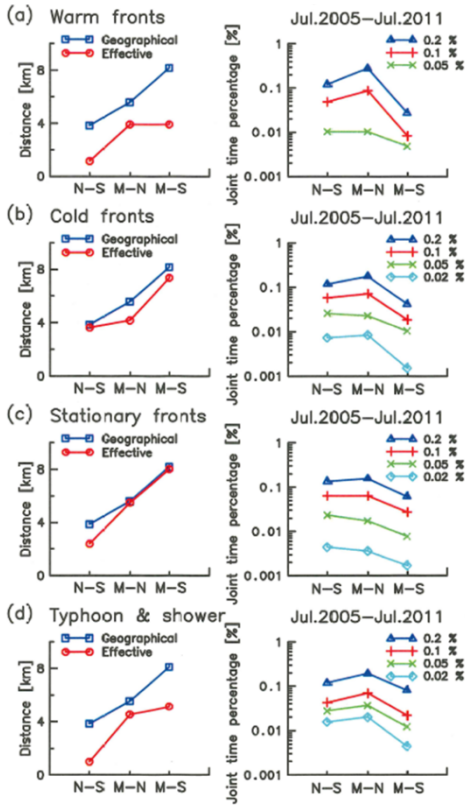


Fig. 11 Site diversity effects of each rain type for the cumulative time percentages of 0.2–0.02% obtained in the narrow area of 3–8 km.

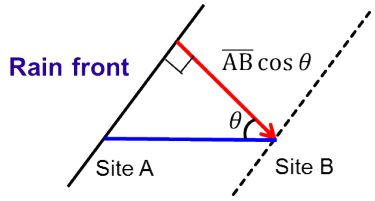


Fig. 10 Relationship between the geographical distance (dark blue line) and the effective distance from site A to site B (red arrow) projected to the passage direction of rain front (black line).

original geographical distances (dark blue). Figure 10 illustrates the relationship between the geographical distance (dark blue line) and the effective distance from site A to site B (red arrow) projected to the passage direction of rain front (black line). The angle θ is here determined by the difference between the geographical alignment of each site shown in Fig. 1 and the average passage directions of the four rain types listed in Tables 2 and 3. Thus, the average lengths projected to the rain area motion are calculated for each rain type. Also, this length $\overline{AB}\cos\theta$ is considered to represent the "effective distance" for the actual site diversity effects between the sites A and B.

As shown in Fig. 10, the effective distance (red) is significantly different according to the average passage angle of each rain type. Therefore, the site diversity effects in Fig. 9 are not necessarily improved by the geographical distances (dark blue) but rather by the effective distance (red). Specifically, in the case of (b) the cold fronts, the effective distance (red) between the sites R-M is slightly longer than the sites O-M. Consequently, the joint time percentages of R-M become smaller and indicate the better site diversity performance than O-M that has the longer geographical distance (dark blue). In the case of (d) typhoon and shower, in contrast, the effective distance (red) between the sites R-M is the shortest, and its time percentages become the largest, indicating the worst performance, regardless of the medium geographical distance (dark blue).

On the other hand, in the case of (c) the stationary fronts along which the rain areas move from nearly west to east as was shown in Fig. 5(a), the effective distances (red) are much the same as the geographical distances (dark blue). Consequently, the joint time percentages are similarly decreased with the geographical distance. In the case of (a) the warm fronts, however, the consistent characteristics of the site diversity effects are not clear due to the lack of data with large attenuation.

In addition, Fig. 11 shows the distances of all pairs of three sites (N-S, M-N, M-S) in the narrow area and the time percentages of the site diversity effects for the rain types of (a) warm, (b) cold, and (c) stationary fronts as well as (d) typhoon and shower, respectively. Their joint time percentages are similarly calculated for the attenuation of 2.2, 3.0, 4.4, and 7.7 dB in the narrow area, which are equivalent to the yearly time percentages of 0.2, 0.1, 0.05, and 0.02% for

the single site, respectively. The attenuation is also selected from the lowest value among the three sites for the single site attenuation similarly to Fig. 8 except for 7.7 dB taken from Moriguchi station. As for (a) the warm front, however, the site diversity effects of the time percentage of 0.02% are not obtained, because the number of data is very sparse.

In the case of the narrow area, the effective distances (red), as a whole, show the same tendency as the geographical distances (dark blue) between the two sites for each rain type. This may be partly because the direction of cold fronts (108.08°) accompanied by large attenuation approaches eastward in Table 3, compared with that of the wide area (132.49°) in Table 2. So, the joint time percentages are basically decreased as the geographical distance (dark blue) is increased, except that the joint time percentages of M-S (Mori-Shijo) are slightly increased in (d) typhoon and shower because of the decrease in the effective distance (red), compared with those in (b) cold and (c) stationary fronts. The same tendency is seen in the joint time percentage of M-S for 0.05% in (a) the warm fronts. Note that the joint time percentages of M-N (Mori-Neya) are, as a whole, increased because the alignment of the two sites nearly coincides with the satellite azimuth angle [2] as was mentioned in Fig. 8.

6. Proposal of Novel Site Diversity Techniques

As was discussed in the previous chapter, site diversity effects are found to largely depend on the effective distance that is defined by the length projected to the direction of rain area motion peculiar to each rain type, such as warm, cold, and stationary fronts or typhoon and shower. Based on these observational results, the direction of rain area motion is examined for each rainfall event, and the pair of the three sites that gives the longest effective distance among them is selected for each passage direction in the wide and narrow areas.

Figure 12 shows the locations of three sites and the relation to the passage direction of rain area motion in the (a) wide area and (b) narrow area, respectively. The 1: blue, 2: green, and 3: red arrows indicate the passage direction of rain area motion that gives the longest effective distance among the pairs of three sites numbered by 1–3 with the same color, respectively. The number of the pair corresponds to 1: OECU-RISH (blue), 2: RISH-MU (green), and 3: OECU-MU (red) in the (a) wide area, while it corresponds to 1: Mori-Neya (blue) 2: Neya-Shijo (green), and 3: Mori-Shijo (red) in the (b) narrow area, respectively.

Also, the broken circular arcs with arrows indicate the range of angles for the passage directions of rain area motion that gives the longest effective distance among the pairs of three sites, as noted in Fig. 12. Table 4 summarizes the relationship between the pair of two sites and the range of these angles.

Here, we propose novel site diversity techniques that select an appropriate pair of two sites from three sites at each rainfall event in real time. In this process, we need to trace

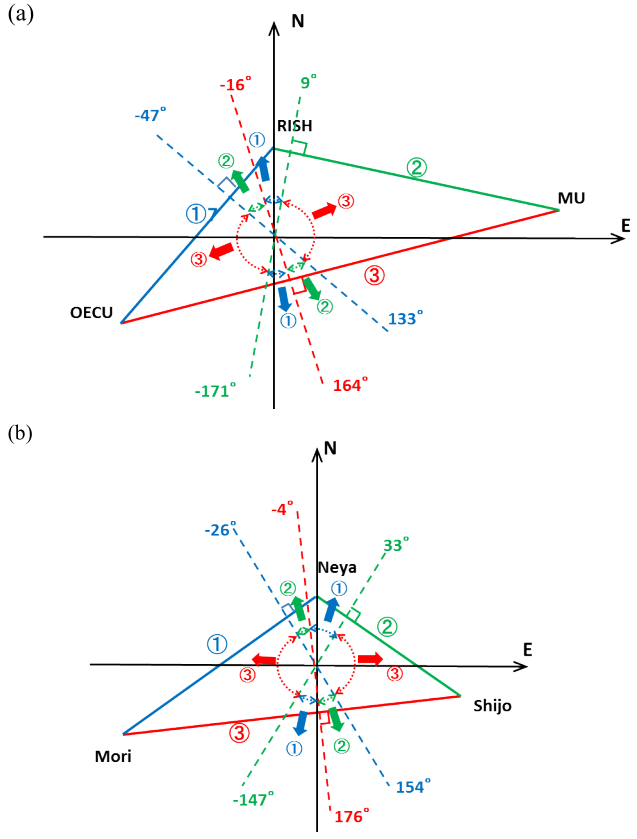


Fig. 12 Locations of three sites and the relation to the passage direction of rain area motion that gives the longest effective distance among the pairs of three sites numbered by 1–3 in the (a) wide and (b) narrow area.

Table 4 Pair of two sites and range of angles for the longest effective distance.

Pair of two sites		Range of angles
Wide area	1: OCEU-RISH (blue)	−16 ~ 9°, 165 ~ 189(−171)°
	2: RISH-MU (green)	−47 ~ − 16°, 133 ~ 164°
	3: OECU-MU (red)	9 ~ 133°, − 171 ~ − 47°
Narrow area	1: Mori-Neya (blue)	−4 ~ 33°, 176 ~ 213(−147)°
	2: Neya-Shijo (green)	−26 ~ − 4°, 154 ~ 176°
	3: Mori-Shijo (red)	33 ~ 154°, − 147 ~ − 26°

the rain area motion at each event to select the two sites in advance. To do this, we consider following three methods.

1. Use average passage directions for each rain types:
   The first method uses a data base such as presented for warm, cold, and stationary fronts, or typhoon and shower during the ten or seven years in this study, as listed in Tables 2 and 3 for the wide and narrow areas, respectively. This method may be time consuming and slightly lack of preciseness, but once obtained, the data base seems to be still useful for wider areas of the Kinki region such as Osaka, Kyoto, Shiga prefectures and so on.
2. Use the motion of fronts or low pressures:
   The second method is nowadays available from weather charts usually published in web sites except for typhoon and shower. Hence, this method seems conve-

nient and practical, although it does not necessarily reflect the rain area motion itself. The correspondence to the real rain area motion is, however, sufficiently shown in Fig. 4, so considerable site diversity effects are expected. The directions of typhoon and shower are, for convenience, assumed to be northward from Tables 2 and 3.

3. Use rain attenuation measurements at three sites:
The third method uses the time difference obtained from cross-correlation functions as shown in Figs. 1 and 2. This is probably the most precise method inferred from the rain attenuation data actually observed by the satellite links that directly reflect the rain area motion, but may be slightly complicated to perform in real time. However, if it is possible to obtain all signals received by the three sites at the same time, this method is the most effective in the practical site diversity operation between two sites instead of switching among three sites in real time.

Based on these considerations, the site diversity techniques using the three proposed methods are compared by numerical calculations in the wide and narrow areas in Figs. 13 and 14, respectively. In these diagrams, the resulting time percentages of the first method switched by rain types (method 1, blue), the second method switched by rain front motions (method 2, dark blue), and the third method switched by rain area motions (method 3, red) are presented, respectively, as well as those of the average of three single sites (black).

Also, the average of simply calculated site diversity effects among two sites (green) is indicated, as well as the site diversity effects among three sites (purple). Here, the simply calculated site diversity effects mean OECU-MU (O-M, red), OECU-RISH (O-R, dark blue), and RISH-MU (R-M, green) in Fig. 6, while these are Mori-Shijo (M-S, red), Mori-Neya (M-N, dark blue) and Neya-Shijo (N-S, green) in Fig. 7.

In the case of the first method, the pairs of two sites are selected for each rain type, in advance, comparing its average passage direction listed in Tables 2 and 3 with the range of angles for the maximum effective distance in Table 4. The results are summarized in Table 5. Specifically, the passage directions of warm front (13.47°), cold front (132.49°), and stationary front (97.55°) in the wide area are included in the range of angles (9∼133°) for the maximum effective distance in the case of 3 (OECU-MU). The direction of typhoon and shower (0.46°) is, however, in the range of those (−16∼9°) for 1 (OECU-RISH) in Table 4. On the other hand, the passage directions of warm front (16.06°) and typhoon and shower (26.68°) in the narrow area are included in the range of angles (−14∼33°) in the case of 1 (Mori-Neya), while cold front (108.08°), and stationary front (71.28°) are in the range of those (33∼154°) for 3 (Mori-Shiho) in Table 4.

In both wide and narrow areas, the time percentages are decreased in the order of the first, second, and third method,

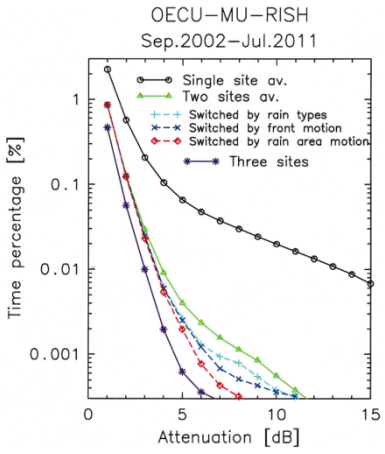


Fig. 13 Time percentages of the proposed three site diversity techniques, together with those of single, two, and three sites, in the wide area.

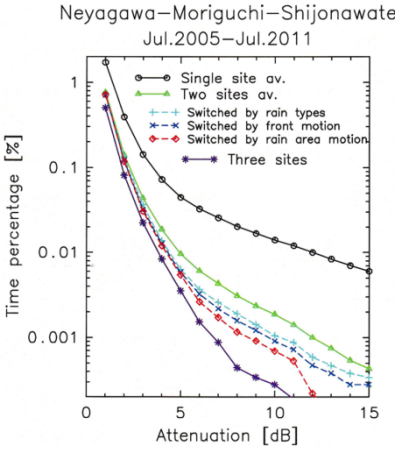


Fig. 14 Time percentages of the proposed three site diversity techniques, together with those of single, two, and three sites, in the narrow area.

Table 5 Pairs of two sites selected for each rain type.

Area	Rain type	Passage direction	Pair of two sites
Wide area	Warm front	13.47 °	3: OECU-MU
	Cold front	132.49 °	3: OECU-MU
	Stationary front	97.55 °	3: OECU-MU
	Typhoon and Shower	0.46 °	1: OECU-RISH
Narrow area	Warm front	16.06 °	1: Mori-Neya
	Cold front	108.08 °	3: Mori-Shijo
	Stationary front	71.28 °	3: Mori-Shijo
	Typhoon and Shower	26.68 °	1: Mori-Neya

so the site diversity effects are further improved in this order. Also, the site diversity effects are improved more than those simply switched between two sites among the three sites (green), and tend to approach those switched among the three sites (purple).

7. Concept of Substantial Distance

Next, the improvement of the site diversity effects due to the three proposed switching methods are presented against the geographical distance d, and compared with the ITU-R

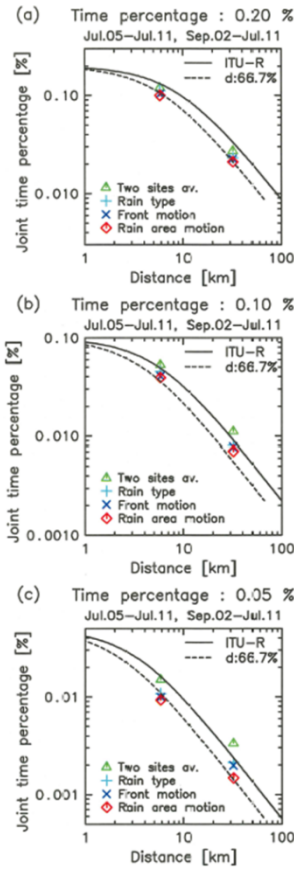


Fig. 15 Joint time percentages of (a) 0.2, (b) 0.1, and (c) 0.05 % for those averaged over two sites (green), switched by rain types (blue), rain front motions (dark blue), and rain area motions (red) [6].

predictions again in Fig. 15 [6]. The distance between the pair of two sites are averaged over the three pairs, and it becomes 5.85 and 32.26 km in the narrow and wide areas, respectively. The joint time percentages P_2 for the original time percentages P_1 of (a) 0.2, and (b) 0.1, and (c) 0.05% are indicated for the following cases: simply averaged two sites (green), switched by rain types (method 1, blue), switched by rain front motions (method 2, dark blue), and switched by rain area motions (method 3, red), respectively, in the same way as was shown in Figs. 13 and 14.

Thin lines indicate joint time percentages P_2 predicted by the ITU-R recommendations for the corresponding time percentages P_1. On the other hand, the dashed lines are distance d' which is required to achieve the same joint time percentage as the original geographical distance d in the ITU-R predictions, when the site diversity is conducted between the two of the three sites. Thus, we define the concept of the "substantial distance" d' that gives the same site diversity effects using the two of the three sites as the conventional ITU-R predictions using the only two sites.

As was shown in Fig. 10, effective distance l for the site diversity between sites A and B should be the length projected to the direction of rain area motion:

$$l = d\cos\theta \tag{3}$$

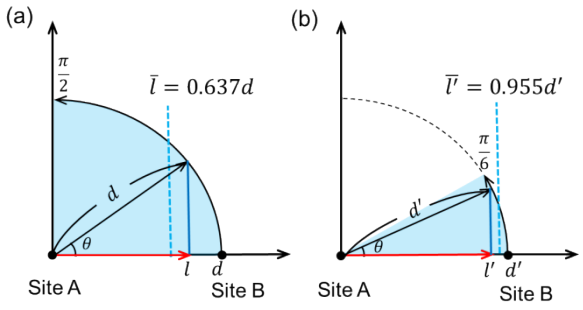


Fig. 16 Relationship between the geographical distance d and the average of effective path length $\bar{l}$ for the passage directions of (a) $\pm\frac{\pi}{2}$ (90°) and (b) $\pm\frac{\pi}{6}$ (30°).

where d is the geographical distance $\overline{AB}$ in Fig. 10. It should be noted here that in Fig. 5, the passage direction of rain area motion is primarily from west to east, and widely distributed from 0° to 180°, although each rain type considerably shows a specific feature of the passage direction. Hence, the ratio of average effective distance to the geographical distance d is given by

$$\overline{\left(\frac{l}{d}\right)} = \int_0^{\frac{\pi}{2}} \cos\theta d\theta \Big/ \left(\frac{\pi}{2}\right) = 0.637 \tag{4}$$

where the passage direction of rain area are considered to be nearly symmetry between northward and southward, so it falls within the range of $\pm\pi/2$ (±90°). Consequently, the integration is performed from 0 to $\pi/2$, using the positive angle side. Thus, we obtain the relationship between the average effective distance $\bar{l}$ and the geographical distance d, when the passage direction of rain area motion is from 0 to 180° as

$$\bar{l} = 0.637d \tag{5}$$

This relation is illustrated in Fig. 16(a).

In the case of using the two sites A and B among the three sites, on the other hand, the passage direction of rain area motion is separated into three sections as shown by Fig. 12, so it falls within the range of $\pm\pi/6$(±30°). Consequently, the integration in Eq. (4) is, on an average, reduced down to $\pi/6$, similarly assuming that the passage direction of rain area is nearly symmetry, although some difference in the angles exists among the pairs of the three sites in Fig. 12. Thus, the ratio of average effective distance to the substantial distance between the two sites d' in this case is given by

$$\overline{\left(\frac{l'}{d'}\right)} = \int_0^{\frac{\pi}{6}} \cos\theta d\theta \Big/ \left(\frac{\pi}{6}\right) = 0.955 \tag{6}$$

From Eq. (6) the relationship between the average effective distance $\bar{l}'$ and the substantial distance d' is given by

$$\bar{l}' = 0.955d' \tag{7}$$

when the passage direction of rain area motion is limited to

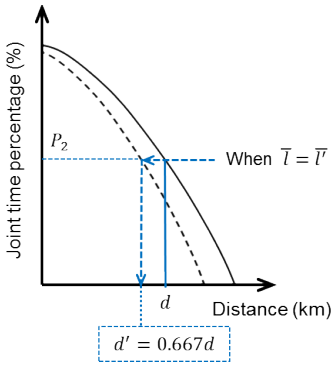


Fig. 17 Relationship between substantial distance d' and original geographical distance d for the same effective distance with a common time percentage P_2, when the site diversity is performed between two of the three sites in the ITU-R predictions.

30° in the case of three sites. This relation is also illustrated in Fig. 16(b).

As was defined earlier, the same joint time percentage P_2 is given by both cases of the original distance d and the substantial distance d', so the average effective distance should be equal between Eqs. (5) and (7) as $\overline{l} = \overline{l'}$. This condition is illustrated in Fig. 17 based on the ITU-R predictions. Thus, the "substantial distance" d' for the site diversity using the two of the three sites is related to the "original geographical distance" d as

$$d' = (0.637/0.955)d = 0.667d \quad (66.7\%) \tag{8}$$

In this sense, the dashed line in Fig. 15 indicates the ITU-R predictions in the case that the substantial distance d' is reduced down to 66.7%. of the original geographical distance d. It is found from Fig. 15 that the joint time percentages switched by rain types (method 1, blue), rain front motions (method 2, dark blue), and rain area motions (method 3, red) gradually approaches to the ITU-R predictions of the substantial distance (dashed line) in this order, while those of averaged two sites (green) stay around the original ITU-R predictions (thin line).

Finally, Fig. 18 illustrates the relationship between the joint time percentages P_2' (dashed line) and P_2 (thin line) obtained by the substantial distance d' and the original geographical distance d in the ITU-R predictions, respectively. Here, P_2' is calculated from the substantial distance d' using Eqs. (1) and (2) by substituting $d = d'/0.667 = 1.499d'$, because the substantial distance is given by $d' = 0.667d$ in Eq. (8). So, this increase of the distances yields the decrease of joint time percentages down to P_2' on the dashed line at the original geographical distances $\overline{\mathrm{AB}}$. The dashed-dotted lines in Fig. 19 show the reduction rates of joint percentages P_2'/P_2 averaged over the original time percentages of 0.2, 0.1, and 0.05% in Fig. 15(a)–(c). Equations (1) and (2) yield the reduction rates of 73.3 and 60.8% for the narrow and wide area three sites, respectively.

These dashed-dotted lines in Fig. 19 are considered to indicate the theoretical limit of site diversity operations choosing the two sites among the three sites in both narrow

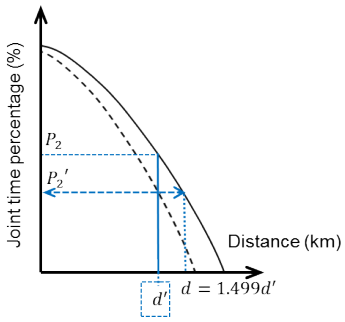


Fig. 18 Relationship between the joint time percentages P_2' (dashed line) and P_2 (thin line) obtained by the substantial distance d' and the original geographical distance d in the ITU-R predictions, respectively.

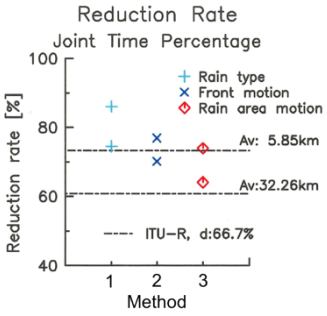


Fig. 19 Comparison of the observed reduction rates of joint time percentages P_2'/P_2 with those obtained from the substantial distance of the site diversity.

and wide areas. In Fig. 19, on the other hand, the observational results on the reduction rate of joint time percentages are obtained from the three kinds of site diversity methods using rain type (method 1, blue), front motions (method 2, dark blue), and rain area motion (method 3, red). Also, they are similarly averaged over the cases of time percentages P_1 of 0.2, 0.1, and 0.05%. Thus in Fig. 19, the observed reduction rates are found to approach these theoretical limits in this order.

As expected from the definition of each site diversity method, the third method (red) based on the rain area motion directly inferred from the satellite links is most effective in the improvement of the site diversity effects. Also, it approaches the theoretical limit of about 61–73% for the reduction in unavailable time percentages, as indicated by dashed-dotted lines in Fig. 19. The first method (blue) using the data base of average passage directions for each rain type, however, seems to still maintain the considerable improvement for the site diversity operations, because it shows the reduction in unavailable time percentages of about 75–85%, compared to the above-mentioned theoretical limit in the distance of 3–50 km.

On the other hand, it should be noted here that if the pair of two sites can be chosen in any direction along the passage of rain area motion, the ratio of average effective distance $\overline{l''}$ to the original two sites is given by

$$\overline{\left(\frac{l''}{\bar{d}}\right)} = \lim_{\theta' \to 0} \int_0^{\theta'} \cos\theta d\theta / \theta' = \lim_{\theta' \to 0} \sin\theta' / \theta' = 1 \qquad (9)$$

This means the average effective distance is equal to the original two sites. In this case, the substantial distance to realize the time percentage given by the ITU-R predictions is reduced down to 63.7% (about 60%) according to Eq. (8). This value may correspond to the theoretical limit of the distance reduction for the site diversity operations using two sites in an arbitrary direction.

8. Conclusions

The improvement of the site diversity effects was discussed in relation to the rain area motions, using the rain attenuation of the Ku-band satellite radio wave signals recently observed at the nearby sites 3–8 km away from each other in Osaka, as well as those obtained in Kyoto and Shiga 20–50 km away from OECU in Osaka. Considering the effective distance between two sites projected to the rain area motions, the substantial distance required to achieve the same site diversity effects are found to be reduced down to about 60% compared to the ITU-R predictions.

Also, this improvement is shown to be realized by novel site diversity techniques choosing two sites with the maximum effective distance projected to rain area motion in the rainfall events obtained on each rainy day from Sept. 2002 to July 2011. Three site diversity methods are newly proposed choosing the pair of sites based on rain type, rain front motion, or rain area motion at each rainfall event. Consequently, the average passage directions of each rain type statistically obtained from long-term observations are found to be useful for practical operations of the site diversity, because the unavailable time percentages are reduced down to about 75–85% compared to the theoretical limit of about 61–73% in the distance of 3–50 km. The more precise method based on the rain area motion is shown to yield the unavailable time percentages near this theoretical limit.

Recently, information of rain area motions is being delivered more easily and precisely in Web sites using advanced radar observations. In future, the site diversity techniques newly presented in this study are expected to be further developed, applying recent image analyzing techniques including such as machine learning using AI techniques to these rain area motions published in Web sites.

References

[1] T. Iida, Satellite Communications, Ohmusha, p.427, 1997 (in Japanese).
[2] ITU-R, P.618-8, "Propagation data and prediction methods required for the design of Earth-space telecommunication systems," ITU-R Recommendations P.618-8, Geneva, 2003.
[3] Y. Maekawa, T. Nakatani, Y. Shibagaki, and T. Hatsuda, "A study on site diversity techniques related to rain area motion using Ku-band satellite signals," IEICE Trans. Commun., vol.E91-B, no.6, pp.1812–1818, June 2008.
[4] Y. Maekawa, K. Sawai, Y. Shibagaki, T. Takami, and K. Hatsuda, "Site diversity effects on Ku-band satellite signal attenuation related to rain area motion," Proc. ISAP2008, 3A07-3, Taipei, Taiwan, Oct. 2008.
[5] Y. Inose and H. Fukuchi, "Analysis of directional dependence of site diversity gain using rain radar data," Proc. ISAP2016, 2A3-2, Okinawa, Japan, Oct. 2016.
[6] Y. Maekawa, "A study on rain attenuation characteristics on propagation paths in satellite links based on long-term observations during thirty years," IEICE Trans. Commun. (Japanese edition), vol.J103-B, no.11, pp.481–490, Nov. 2020.
[7] Y. Inamori, Y. Shibagaki, and Y. Maekawa, "Rain attenuation characteristics of Ku-band satellite signals in relation to the wind velocities observed on the ground," Proc. ISAP 2012, 4D3-4, P0235, Nagoya, Japan, Oct.-Nov. 2012.

Yasuyuki Maekawa received the B.S., M.S., and Ph.D. degrees in electronics engineering from Kyoto University in 1979, 1981, and 1985, respectively. In 1985, he jointed the Department of Tele-communication Engineering, Osaka Electro-Communication University, Neyagawa, Osaka, Japan, where he is currently a Professor of satellite communication engineering. His research activities are concerned with microwave and millimeter wave propagation on the earth-space paths and meteorological radar observations of the atmosphere. Dr. Maekawa is a Member of Institute of Electrical and Electronics Engineers (IEEE), the Society of Geomagnetism and Earth, Planetary and Space Science, and the Meteorological Society of Japan.

Yoshiaki Shibagaki received the B.S., M.S., and Ph.D. degrees in communication engineering from Osaka Electro-Communication University in 1992, 1994, and 1997, respectively. After working as a research fellow PD of the Japan Society for the Promotion of Science (JSPS) in Kyoto University, Uji, Kyoto, he joined Department of Telecommunication Engineering, Osaka Electro-Communication University in 1999. He is currently a Professor of applied radio wave engineering including remote sensing techniques. His main research activities are observations of mesoscale meteorology using atmospheric and meteorological radars. He is a Member of the Meteorological Society of Japan.

Tomoyuki Takami received the B.S. degree in science, the M.S. and Ph.D. degrees in electronics engineering from Kyoto University in 1980, 1988 and 1992, respectively. In 1997, he joined the Department of Information Science and Arts, Osaka Electro-Communication University, Shijonawate, Osaka, Japan, where he is currently a Professor of digital games. His research activities are concerned with game development methodology and radar observations of the ionospheric atmosphere. Dr. Takami is a Member of Game Amusement Society, the Society of Geomagnetism and Earth, Planetary and Space Science, and the Society for Art and Science.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

Joint PAPR Reduction Using Null Space in MIMO Channel and Predistortion for MIMO-OFDM Signals in Multi-Antenna AF-Type Relay Transmission

Asuka KAKEHASHI[†], ***Member and*** **Kenichi HIGUCHI**[†a)], ***Senior Member***

SUMMARY The combination of peak-to-average power ratio (PAPR) reduction and predistortion (PD) techniques effectively reduces the nonlinear distortion of a transmission signal caused by power amplification and improves power efficiency. In this paper, assuming downlink amplify-and-forward (AF)-type relaying of multiple-input multiple-output (MIMO)-orthogonal frequency division multiplexing (OFDM) signals, we propose a joint method that combines a PD technique with our previously reported PAPR reduction method utilizing the null space of a MIMO channel. In the proposed method, the reported PAPR reduction method reduces the PAPR at a relay station (RS) as well as that at a base station (BS) by using only signal processing at the BS. The PD process at the BS and RS further reduces the nonlinear distortion caused by nonlinear power amplification. Computer simulation results show that the proposed method enhances the effectiveness of PD at the BS and RS and achieves further coverage enhancement compared to conventional methods.

key words: *peak-to-average power ratio (PAPR), predistortion, relaying, amplify-and-forward, orthogonal frequency division multiplexing (OFDM), multiple-input multiple-output (MIMO), null space*

1. Introduction

The combination of massive multiple-input multiple-output (MIMO) [1], [2] using beamforming (BF) and orthogonal frequency division multiplexing (OFDM) signals offers wide-coverage enhanced mobile broadband. In the 5th generation mobile communication system (5G) New Radio (NR) [3] and beyond [4], the importance of relay transmission [5]–[7] increases in order to further expand the coverage area in a high frequency band such as the millimeter-wave band. In this paper, we consider downlink MIMO-OFDM transmission using amplify-and-forward (AF)-type relaying. Here, the AF relay station (RS) is equipped with multiple antennas and transfers the received signal from the base station (BS) to a set of user equipment (UE) after power amplification without decoding the data signal.

The high peak-to-average power ratio (PAPR) of OFDM signals is a significant drawback. When a signal with a high PAPR is amplified using a nonlinear power amplifier (PA), severe spectral dispersion and in-band distortion can occur. Increasing the amount of input backoff (IBO) of the PA and the use of a predistortion (PD) technique before power amplification are major approaches to addressing this problem by increasing the linear range of the PA. Increasing the IBO moves the operating point of the PA away from the saturation point and allows it to operate in the linear region. However, the power amplification efficiency is severely reduced. This is a particularly serious problem in massive MIMO due to the limited linearity requirement on the PA for each of the huge number of antennas. PD is applied before power amplification to extend the linear range of the PA. However, it cannot handle input signals above the saturation input power. PAPR reduction is important to address these problems [8]. This is the same for the RS, and PAPR reduction at both the BS and RS transmitters is important to achieve high-speed high-quality transmission with wide coverage for multi-antenna relay transmission.

1.1 Related Work

A number of PAPR reduction methods employing MIMO-OFDM have been investigated, e.g., in [9]–[17]. Among these methods, when a powerful channel code such as the turbo code or low-density parity check (LDPC) code is employed, [18] revealed that a PAPR reduction method that does not reduce the frequency efficiency at the cost of in-band interference such as the clipping and filtering (CF) method [10], [11] is superior to those that consume a part of the frequency bandwidth to reduce the PAPR such as the tone reservation method [14] from the viewpoint of the tradeoff between the PAPR reduction and the error rate for a given data rate. However, the in-band PAPR reduction signal added to the data signal at the transmitter in the CF method becomes a source of interference to the data streams at the receiver.

To address this problem, PAPR reduction methods utilizing the number of antennas at the BS that is sufficiently larger than the number of antennas at the UE receiver in a downlink massive MIMO environment were reported in [19]–[32]. In [19], [20], some of antennas at the BS are used exclusively for transmitting compensation signals that eliminate the PAPR reduction signal. The PAPR reduction signal transmitted from the antenna dedicated to the data signal is canceled on the UE receiver end using the signal transmitted from the antenna exclusively used for transmitting the compensation signals. In [21]–[32], PAPR reduction is performed by utilizing the null space of the downlink

Manuscript received January 17, 2024.
Manuscript revised April 30, 2024.
Manuscript publicized June 28, 2024.
[†]Graduate School of Science and Technology, Tokyo University of Science, Noda-shi, 278-8510 Japan.
a) E-mail: higuchik@rs.tus.ac.jp
DOI: 10.23919/transcom.2024EBP3017

MIMO channel, which has the dimension of the difference between the number of transmitter antennas at the BS and the number of receiver antennas at the UEs. This method restricts the PAPR reduction signal to be transmitted only to the null space in MIMO channels by using BF. This restriction suppresses interference due to PAPR reduction to the data stream at the UE receiver. Unlike the methods in [19], [20], they can use all the transmitter antennas for data transmission and obtain higher BF gains (received signal power gain and interference suppression gain).

In downlink AF-type relay transmission with a single transmitter antenna at the BS, PAPR reduction for the RS transmission signal is not so important because if the PAPR on the BS transmitter end is reduced, the PAPR on the RS transmitter end can also be reduced. There have been various studies on PAPR reduction methods for the OFDM signal in AF-relay transmission with a single transmitter antenna at the BS. PAPR reduction based on CF for the BS transmitter was investigated in [33] and a method combining PAPR reduction based on CF for the BS transmitter and PD was investigated in [34]. Also, PAPR reduction based on a partial transmit sequence (PTS) for the BS transmitter was investigated in [35] and a method combining PAPR reduction based on PTS for the BS transmitter and PD was investigated in [36]. These studies show that PD should be combined with PAPR reduction to increase the effectiveness of PD.

In AF-type relay transmission with multiple transmitter antennas at the BS, the received signal at the RS is a superposition of many transmission signals from the BS, so even if the PAPR at the BS is reduced, the PAPR at the RS will increase again. Therefore, in this case, it is very important to reduce the PAPR not only on the BS transmitter end but also on the RS transmitter end. To the best of our knowledge, there have been few studies on PAPR reduction methods for OFDM signals in AF-relay transmission when the BS has multiple transmitter antennas. Members of our research group reported a PAPR reduction method utilizing the null space in a MIMO channel for a multi-antenna AF-type RS [37], [38]. In this method, PAPR reduction utilizing the null space in the overall MIMO channel, which combines the channel between the BS and RS and that between the RS and the UEs, is applied to the transmission signal at the BS in advance. Then, the peak signal component observed again at the RS receiver due to the effect of the MIMO channel is reduced by generating a PAPR reduction signal at the RS that is transmitted only to the null space in the MIMO channel between the RS and UEs. This method enables PAPR reduction on both the BS and RS transmitter ends while suppressing interference to the data stream at the UE receivers. However, this method requires complex signal processing at the RS for PAPR reduction. If the signal processing delay exceeds the cyclic prefix length of the OFDM signal, relay transmission using the same channel as is used at the BS, which is an advantage of AF-relay transmission, will not be possible to avoid interference from the delayed signal. Therefore, there is concern regarding the reduction in channel capacity (throughput) for a method that needs signal processing for PAPR reduction at the RS such as [37], [38].

Members of our group recently reported a method to achieve PAPR reduction for the RS transmitter by relying only on the signal processing at the BS while utilizing the null space in the MIMO channel [39], [40]. In this method, signal processing at the BS alternately and repeatedly generates PAPR reduction signals for both the BS and RS transmitters. These PAPR reduction signals are projected onto the null space of the integrated MIMO channel of the entire system. After iterative processing, these PAPR reduction signals are transmitted together from the BS. This method reduces not only the PAPR at the BS but also the PAPR at the RS and achieves higher throughput without potential channel capacity degradation caused by the delay due to the PAPR reduction signal processing at the RS.

1.2 Contributions and Organization

When joint PAPR reduction and PD, e.g., in [34] and [36], are applied to AF-type relay transmission with multiple transmitter antennas at the BS, a high PAPR at the RS becomes a problem, which weakens the effectiveness of the PD at the RS and increases the IBO of the PA at the RS. To address this problem, this paper proposes a method combining the previously reported PAPR reduction method in [40] with PD. The proposed method uses the previously reported PAPR reduction method and PD processing for the data signals at the BS. In addition, PD is applied to the received signals at the RS. Since the previous PAPR reduction method achieves PAPR reduction both for the BS and RS transmitters without signal processing at the RS while the processing delay of PD is in general very short, the proposed method enhances the effectiveness of PD at both the BS and RS and improves the transmission performance of the downlink multi-antenna AF relaying without causing a prohibitive processing delay at the RS. The technical difference between the proposed method and [40] is that, in addition to performing the previously reported PAPR reduction, PD processing is performed before power amplification at the BS and RS. The previous PAPR reduction method in [40] enhances the effectiveness of PD at both the BS and RS thanks to the reduced peak signal power of the transmission signal to be PD processed, resulting in throughput improvement.

In this paper, the method in [34], which combines CF and PD with AF-type MIMO-OFDM relay transmission with multiple transmitter antennas, is assumed as the conventional method to facilitate the analysis of the difference between the proposed method and the previous method. The advantages of the proposed method over the conventional method are the suppression of the interference to the data stream at the UE receiver by utilizing the null space of the MIMO channel, and the reduction of the PAPR at the RS by using the signal processing at the BS, which enhances the effectiveness of the PD at the RS. Computer simulations

show that the proposed method achieves higher throughput than that for the conventional method.

The remainder of the paper is organized as follows. First, Sect. 2 describes the system model. Section 3 presents the proposed method. Section 4 shows the numerical results based on computer simulations. Finally, Sect. 5 concludes the paper.

2. System Model

Figure 1 shows the system model assumed in this paper. The distance between the BS and RS, that between the RS and UEs, and that between the BS and UEs are denoted as D_{BR}, D_{RU}, and D_{BU}, respectively. The number of BS transmission antennas is N_{B} and those for RS transmission and reception are N_{R}. We consider a downlink multiuser MIMO scenario where N_{U} users each having a single receiver antenna are spatially multiplexed. Assuming a massive MIMO environment, we set $N_{\mathrm{B}} \geq N_{\mathrm{R}} > N_{\mathrm{U}}$. The total number of subcarriers in the OFDM signal is K. There are B frequency blocks under different channel conditions from each other. The number of subcarriers in each frequency block is K/B. For simplicity, the channel variation within a frequency block is omitted. At frequency block b ($b = 1, \ldots, B$), the $N_{\mathrm{R}} \times N_{\mathrm{B}}$-dimensional channel matrix between the BS and RS is denoted as $\mathbf{H}_{\mathrm{BR},b}$ and the $N_{\mathrm{U}} \times N_{\mathrm{R}}$-dimensional channel matrix between the RS and UEs is denoted as $\mathbf{H}_{\mathrm{RU},b}$. The $N_{\mathrm{U}} \times N_{\mathrm{B}}$-dimensional channel matrix of the direct link between the BS and UEs is denoted as $\mathbf{H}_{\mathrm{BU},b}$. The $N_{\mathrm{U}} \times N_{\mathrm{B}}$-dimensional channel matrix of the overall relay link is denoted as $\mathbf{H}_{\mathrm{BRU},b} = A_{\mathrm{R}}\mathbf{H}_{\mathrm{RU},b}\mathbf{H}_{\mathrm{BR},b}$ where A_{R} is the power amplification gain in the RS. The overall MIMO channel that combines the channels of the direct and relay links is denoted as $\mathbf{H}_{\mathrm{T},b} = \mathbf{H}_{\mathrm{BRU},b} + \mathbf{H}_{\mathrm{BU},b}$. It is assumed that the BS knows all channel matrixes in advance.

The N_{U}-dimensional data stream vector before BF on subcarrier k ($k = 1, \ldots, K/B$) of each frequency block b is denoted as $\mathbf{s}_{b,k}$. The BS generates the N_{B}-dimensional transmission signal vector after BF, $\mathbf{x}_{b,k}$, by multiplying $\mathbf{s}_{b,k}$ by $N_{\mathrm{B}} \times N_{\mathrm{U}}$-dimensional BF matrix $\mathbf{B}_b$. The proposed method presented in Sect. 3 can be applied to any BF method for data signals. In this paper, $\mathbf{B}_b$ is generated based on zero forcing (ZF) for integrated channel matrix $\mathbf{H}_{\mathrm{T},b}$. Thus, $\mathbf{x}_{b,k}$ is represented as

$$\begin{aligned}\mathbf{x}_{b,k} &= \mathbf{B}_b\mathbf{s}_{b,k} = \mathbf{H}_{\mathrm{T},b}^{-}\mathbf{s}_{b,k} = \left(\mathbf{H}_{\mathrm{BRU},b} + \mathrm{H}_{\mathrm{BU},b}\right)^{-}\mathbf{s}_{b,k} \\ &= \left(A_{\mathrm{R}}\mathbf{H}_{\mathrm{RU},b}\mathbf{H}_{\mathrm{BR},b} + \mathbf{H}_{\mathrm{BU},b}\right)^{-}\mathbf{s}_{b,k}\end{aligned}, \quad (1)$$

$\mathbf{H}_{\mathrm{T},b}^{-}$ is a Moore-Penrose generalized inverse matrix of $\mathbf{H}_{\mathrm{T},b}$ ($\mathbf{H}_{\mathrm{T},b}^{-} = \mathbf{H}_{\mathrm{T},b}^{H}(\mathbf{H}_{\mathrm{T},b}\mathbf{H}_{\mathrm{T},b}^{H})^{-1}$).

The $N_{\mathrm{U}} \times K/B$-dimensional data stream matrix in frequency block b is denoted as $\mathbf{S}_b = \begin{bmatrix}\mathbf{s}_{b,1} & \cdots & \mathbf{s}_{b,K/B}\end{bmatrix}$. The $N_{\mathrm{B}} \times K/B$-dimensional transmission data signal matrix after BF, $\mathbf{X}_b$, in frequency block b is expressed as

$$\begin{aligned}\mathbf{X}_b &= \mathbf{B}_b\mathbf{S}_b = \mathbf{H}_{\mathrm{T},b}^{-}\mathbf{S}_b = \left(\mathbf{H}_{\mathrm{BRU},b} + \mathbf{H}_{\mathrm{BU},b}\right)^{-}\mathbf{S}_b \\ &= \left(A_{\mathrm{R}}\mathbf{H}_{\mathrm{RU},b}\mathbf{H}_{\mathrm{BU},b} + \mathbf{H}_{\mathrm{BU},b}\right)^{-}\mathbf{S}_b\end{aligned}. \quad (2)$$

The transmission signal matrix at the BS after the PAPR reduction process is expressed as $\mathbf{X}_b + \mathbf{E}_b$, where $\mathbf{E}_b$ is the PAPR reduction signal matrix generated at the BS. PD processing and power amplification are performed on $\mathbf{X}_b + \mathbf{E}_b$. If the PD process at the BS is ideal, $\mathbf{X}_b + \mathbf{E}_b$ is linearly amplified, which is expressed as $A_{\mathrm{B}}(\mathbf{X}_b + \mathbf{E}_b)$, where A_{B} represents a power amplification gain in the BS.

In this case, the $N_{\mathrm{R}} \times K/B$-dimensional received signal matrix at the RS, $\mathbf{Y}_{\mathrm{R},b}$, is represented as

$$\mathbf{Y}_{\mathrm{R},b} = \mathbf{H}_{\mathrm{BR},b}A_{\mathrm{B}}\left(\mathbf{X}_b + \mathbf{E}_b\right) + \mathbf{Z}_{\mathrm{R},b}, \quad (3)$$

where $\mathbf{Z}_{\mathrm{R},b}$ is the noise matrix observed at the RS receiver in frequency block b. At the RS, PD processing and power amplification are performed on $\mathbf{Y}_{\mathrm{R},b}$. If we assume ideal PD at the RS for simplicity, $\mathbf{Y}_{\mathrm{R},b}$ is linearly amplified, which is expressed as $A_{\mathrm{R}}\mathbf{Y}_{\mathrm{R},b}$.

The $N_{\mathrm{U}} \times K/B$-dimensional received signal matrix, $\mathbf{Y}_{\mathrm{U},b}$, at N_{U} UEs is expressed as the following.

$$\begin{aligned}\mathbf{Y}_{\mathrm{U},b} &= \mathbf{H}_{\mathrm{RU},b}A_{\mathrm{R}}\mathbf{Y}_{\mathrm{R},b} + \mathbf{H}_{\mathrm{BU},b}A_{\mathrm{B}}\left(\mathbf{X}_b + \mathbf{E}_b\right) + \mathbf{Z}_{\mathrm{U},b} \\ &= \mathbf{H}_{\mathrm{RU},b}A_{\mathrm{R}}\mathbf{H}_{\mathrm{BR},b}A_{\mathrm{B}}\left(\mathbf{X}_b + \mathbf{E}_b\right) + \mathbf{H}_{\mathrm{RU},b}A_{\mathrm{R}}\mathbf{Z}_{\mathrm{R},b} \\ &\quad + \mathbf{H}_{\mathrm{BU},b}A_{\mathrm{B}}\left(\mathbf{X}_b + \mathbf{E}_b\right) + \mathbf{Z}_{\mathrm{U},b} \\ &= A_{\mathrm{B}}\mathbf{H}_{\mathrm{T},b}\left(\mathbf{X}_b + \mathbf{E}_b\right) + A_{\mathrm{R}}\mathbf{H}_{\mathrm{RU},b}\mathbf{Z}_{\mathrm{R},b} + \mathbf{Z}_{\mathrm{U},b} \\ &= A_{\mathrm{B}}\mathbf{S}_b + A_{\mathrm{B}}\mathbf{H}_{\mathrm{T},b}\mathbf{E}_b + A_{\mathrm{R}}\mathbf{H}_{\mathrm{RU},b}\mathbf{Z}_{\mathrm{R},b} + \mathbf{Z}_{\mathrm{U},b}\end{aligned}, \quad (4)$$

where $\mathbf{Z}_{\mathrm{U},b}$ is the noise matrix observed at the UE receivers in frequency block b.

The explanation here assumes ideal PD to simplify the description, where the input signals to the PD have amplitudes that are always lower than the saturation level of the PD. However, in a real system, the IBO is reduced to the extent that it satisfies the required adjacent channel leakage ratio (ACLR). Therefore, the signal components with amplitudes higher than the saturation level, which the PD cannot address, are also input to the PD process. This results in nonlinear distortion through power amplification. To suppress this nonlinear distortion, it is important to reduce the PAPR both on the BS and RS transmitter ends using PAPR reduction signal $\mathbf{E}_b$ to reduce signal components whose amplitudes are higher than the saturation value.

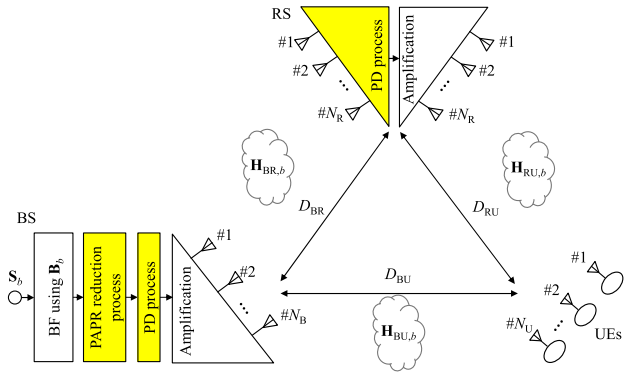


Fig. 1 System model.

3. Proposed Method

In the proposed method, the PAPR reduction method using the null space in the MIMO channel [40] is first applied at the BS as shown in Fig. 1. After that, PD is processed at both the BS and RS. The purpose of the proposed method is to reduce the PAPR of the signal as much as possible before PD at the BS and RS while suppressing interference component $A_{\mathrm{B}}\mathbf{H}_{T,b}\mathbf{E}_b$ in (4) due to the PAPR reduction signal. This is accomplished by using the PAPR reduction method that utilizes the null space in the MIMO channel, and thereby enhances the effectiveness of the PD at the BS and RS.

3.1 PAPR Reduction Using Null Space in MIMO Channel Before PD

The PAPR reduction process alternately repeats generating the PAPR reduction signal that reduces the PAPR at the BS transmitter and the PAPR reduction signal that reduces the PAPR at the RS transmitter as shown in Fig. 2. The PAPR reduction signal matrix in frequency block b that reduces the PAPR at the RS transmitter generated at the j-th ($j = 1, \ldots, J$: J is the number of iterations) iteration is denoted as $\mathbf{E}^{(j)}_{\mathrm{R},b}$. Similarly, the PAPR reduction signal matrix in frequency block b that reduces the PAPR at the BS transmitter generated at the j-th iteration is denoted as $\mathbf{E}^{(j)}_{\mathrm{B},b}$. As the initial setting, $\mathbf{E}^{(0)}_{\mathrm{R},b}$ and $\mathbf{E}^{(0)}_{\mathrm{B},b}$ are set to zero.

In the following, we explain the proposed method focusing on signal processing in frequency block b. In the proposed method, the same processing is performed for all frequency blocks. At the j-th iteration, $\mathbf{E}^{(j)}_{\mathrm{R},b}$ is generated first. For the temporally assumed transmission signal matrix of the BS at the j-th iteration, $\mathbf{X}_f + \mathbf{E}^{(j-1)}_{\mathrm{R},b} + \mathbf{E}^{(j-1)}_{\mathrm{B},b}$, the received signal matrix, $\tilde{\mathbf{Y}}^{(j)}_{\mathrm{R},b}$, at the RS is estimated as

$$\tilde{\mathbf{Y}}^{(j)}_{\mathrm{R},b} = \mathbf{H}_{\mathrm{B},b}\left(\mathbf{X}_b + \mathbf{E}^{(j-1)}_{\mathrm{R},b} + \mathbf{E}^{(j-1)}_{\mathrm{B},b}\right). \tag{5}$$

The K-dimensional time domain received signal vector, $\tilde{\mathbf{y}}^{(j)}_{\mathrm{R},n}$, at antenna n ($n = 1, \ldots, N_{\mathrm{R}}$) of the RS is represented as

$$\left[\tilde{\mathbf{y}}^{(j)}_{\mathrm{R},1} \quad \cdots \quad \tilde{\mathbf{y}}^{(j)}_{\mathrm{R},N_{\mathrm{R}}}\right] = \mathbf{F}^H\left[\tilde{\mathbf{Y}}^{(j)}_{\mathrm{R},1} \quad \cdots \quad \tilde{\mathbf{Y}}^{(j)}_{\mathrm{R},B}\right]^T. \tag{6}$$

Here, $\mathbf{F}$ is the $K \times K$-dimensional fast Fourier transform (FFT) matrix and $\mathbf{F}^H$ is the inverse FFT (IFFT) matrix. The time-domain PAPR reduction signal vector, $\tilde{\boldsymbol{\delta}}^{(j)}_{\mathrm{R},n}$, at antenna n of the RS is obtained by applying CF to $\tilde{\mathbf{y}}^{(j)}_{\mathrm{R},n}$ in order to reduce the PAPR. The $N_{\mathrm{R}} \times K/B$-dimensional frequency-domain PAPR reduction signal matrix in frequency block b, $\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$, is represented as

$$\left[\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},1} \quad \cdots \quad \tilde{\mathbf{E}}^{(j)}_{\mathrm{R},B}\right] = \left[\tilde{\boldsymbol{\delta}}^{(j)}_{\mathrm{R},1} \quad \cdots \quad \tilde{\boldsymbol{\delta}}^{(j)}_{\mathrm{R},N_{\mathrm{R}}}\right]^T \mathbf{F}^T. \tag{7}$$

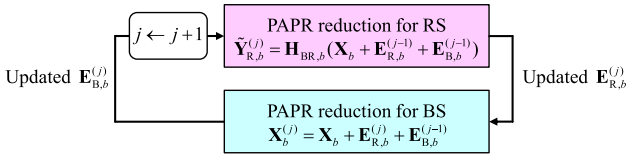


Fig. 2 Iterative PAPR reduction process in proposed method.

However, all the PAPR reduction signals are transmitted from the BS in the proposed method. The PAPR reduction signal matrix for the RS that needs to be transmitted from the BS is represented as $\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$, which is received at the RS as $\mathbf{H}_{\mathrm{BR,b}}\left(\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}\right) = \tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$, where $\mathbf{H}^{-}_{\mathrm{BR},b}$ is the Moore-Penrose generalized inverse matrix of $\mathbf{H}_{\mathrm{BR},b}$. However, $\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$ contains components that interfere with the data stream at the UE receivers. To remove the interference observed at the UE receivers, $\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$ is projected onto the null space in overall MIMO channel $\mathbf{H}_{\mathrm{T},b}$. Since $\mathbf{H}_{\mathrm{T},b}$ is the $N_{\mathrm{U}} \times N_{\mathrm{B}}$-dimensional matrix and we assume $N_{\mathrm{B}} > N_{\mathrm{U}}$, we have $N_{\mathrm{B}} \times (N_{\mathrm{B}} - N_{\mathrm{U}})$-dimensional matrix $\mathbf{V}_{\mathrm{T},b}$ corresponding to the null space in $\mathbf{H}_{\mathrm{T},b}$. Thus, $\mathbf{H}_{\mathrm{T},b}\mathbf{V}_{\mathrm{T},b} = \mathbf{O}$ and we assume that all column vectors of $\mathbf{V}_{T,b}$ are orthonormal. The projection of $\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{R,b}$ onto $\mathbf{V}_{\mathrm{T},b}$ is represented as $\mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$. By adding $\mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}$ to $\mathbf{E}^{(j-1)}_{\mathrm{R},b}$, $\mathbf{E}^{(j-1)}_{\mathrm{R},b}$ is updated to $\mathbf{E}^{(j)}_{\mathrm{R},b}$ as

$$\mathbf{E}^{(j)}_{\mathrm{R},b} = \mathbf{E}^{(j-1)}_{\mathrm{R},b} + \mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\mathbf{H}^{-}_{\mathrm{BR},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{R},b}. \tag{8}$$

Next, $\mathbf{E}^{(j)}_{\mathrm{B},b}$ is generated. PAPR reduction utilizing null space $\mathbf{V}_{\mathrm{T},b}$ in $\mathbf{H}_{\mathrm{T},b}$ is performed on the temporally assumed transmission signal matrix, $\mathbf{X}^{(j)}_b = \mathbf{X}_b + \mathbf{E}^{(j)}_{\mathrm{R},b} + \mathbf{E}^{(j-1)}_{\mathrm{B},b}$, at the BS. The K-dimensional time domain transmission signal vector, $\mathbf{y}^{(j)}_{\mathrm{B},n}$, at antenna n is represented as

$$\left[\mathbf{y}^{(j)}_{\mathrm{B},1} \quad \cdots \quad \mathbf{y}^{(j)}_{\mathrm{B},N_{\mathrm{B}}}\right] = \mathbf{F}^H\left[\mathbf{X}^{(j)}_1 \quad \cdots \quad \mathbf{X}^{(j)}_B\right]^T. \tag{9}$$

The time-domain PAPR reduction signal vector, $\tilde{\boldsymbol{\delta}}^{(j)}_{\mathrm{B},n}$, at antenna n of the BS is obtained by applying CF to $\mathbf{y}^{(j)}_{\mathrm{B},n}$ in order to reduce the PAPR. The $N_{\mathrm{B}} \times K/B$-dimensional frequency domain PAPR reduction signal matrix in frequency block b, $\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}$, is represented as

$$\left[\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},1} \quad \cdots \quad \tilde{\mathbf{E}}^{(j)}_{\mathrm{B},B}\right] = \left[\boldsymbol{\delta}^{(j)}_{\mathrm{B},1} \quad \cdots \quad \boldsymbol{\delta}^{(j)}_{\mathrm{B},N_{\mathrm{B}}}\right]^T \mathbf{F}^T. \tag{10}$$

To remove the interference observed at the UE receivers, $\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}$ is projected onto $\mathbf{V}_{\mathrm{T},b}$. The projection of $\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}$ onto $\mathbf{V}_{T,b}$ is represented as $\mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}$. By adding $\mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}$ to $\mathbf{E}^{(j-1)}_{\mathrm{B},b}$, $\mathbf{E}^{(j-1)}_{\mathrm{B},b}$ is updated to $\mathbf{E}^{(j)}_{\mathrm{B},b}$ as

$$\mathbf{E}^{(j)}_{\mathrm{B},b} = \mathbf{E}^{(j-1)}_{\mathrm{B},b} + \mathbf{V}_{\mathrm{T},b}\mathbf{V}^H_{\mathrm{T},b}\tilde{\mathbf{E}}^{(j)}_{\mathrm{B},b}. \tag{11}$$

The above process is repeated J times, and the BS finally transmits $\mathbf{X}^{(J)}_b = \mathbf{X}_b + \mathbf{E}^{(J)}_{\mathrm{R},b} + \mathbf{E}^{(J)}_{\mathrm{B},b}$.

In the proposed method, interference component $\mathbf{H}_{\mathrm{T},b}\mathbf{E}_b = \mathbf{H}_{\mathrm{T},b}(\mathbf{E}^{(J)}_{\mathrm{R},b} + \mathbf{E}^{(J)}_{\mathrm{B},b})$ due to the PAPR reduction signal in $\mathbf{Y}_{\mathrm{U},b}$ in (4) is zero in principle. Therefore, the proposed method reduces the PAPR at the RS transmitter in addition

to that at the BS transmitter by using only the signal processing at the BS, while suppressing the interference to the data stream at the UE receivers.

3.2 PD Process

In the proposed method, after the PAPR reduction process described in Sect. 3.1, the PD is processed both at the BS and RS. In this paper, the Rapp model [41] is assumed as the solid state power amplifier (SSPA) model since the Rapp Model is a widely accepted SSPA model. The amplitude modulation-amplitude modulation (AM-AM) transfer function of the Rapp SSPA model is defined as

$$x_{\mathrm{PA,out}}(t) = G_0 \frac{x_{\mathrm{PA,in}}(t)}{\left(1 + \left(\frac{\left|x_{\mathrm{PA,in}}(t)\right|}{A_{\mathrm{SAT}}}\right)^{2p}\right)^{1/2p}}. \tag{12}$$

The input and output signals of the SSPA are respectively denoted as $x_{\mathrm{PA,in}}(t)$ and $x_{\mathrm{PA,out}}(t)$ at time t. Term G_0 is the amplification gain in the linear domain of the amplifier and A_{SAT} is the saturation value of the input amplitude. Term p is the parameter controlling the smoothness of the transition from the linear region to the saturation region. Meanwhile, the amplitude modulation-phase modulation (AM-PM) transfer function of the Rapp SSPA model is ideal, thus the output phase change is assumed to be always zero.

PD is performed before power amplification to compensate for non-linear distortion in the PA. The PD of the Rapp SSPA model is the inverse function of (12) with G_0 removal and is expressed as

$$x_{\mathrm{PD,out}}(t) = \begin{cases} \dfrac{x_{\mathrm{PD,in}}(t)}{\left(1 - \left(\frac{\left|x_{\mathrm{PD,in}}(t)\right|}{A_{\mathrm{SAT}}}\right)^{2p}\right)^{1/2p}}, & \left|x_{\mathrm{PD,in}}(t)\right| < A_{\mathrm{SAT}} \\ x_{\mathrm{PD,in}}(t), & \left|x_{\mathrm{PD,in}}(t)\right| \geq A_{\mathrm{SAT}} \end{cases}. \tag{13}$$

In (13), the input and output signals of PD are respectively denoted as $x_{\mathrm{PD,in}}(t)$ and $x_{\mathrm{PD,out}}(t)$ at time t. The output signal of PD, $x_{\mathrm{PD,out}}(t)$, corresponds to the input signal of the SSPA, $x_{\mathrm{PA,in}}(t)$. Substituting (13) into (12), we obtain $x_{\mathrm{PA,out}}(t) = G_0 x_{\mathrm{PD,in}}(t)$, confirming that the output signal of the PA is linearized when the amplitude of $x_{\mathrm{PD,in}}(t)$, $|x_{\mathrm{PD,in}}(t)|$, is lower than A_{SAT}.

PD does not work on signals whose input amplitude $|x_{\mathrm{PD,in}}(t)|$ is equal to or greater than A_{SAT}, resulting in $x_{\mathrm{PD,out}}(t) = x_{\mathrm{PD,in}}(t)$. Therefore, it is important to reduce the PAPR of the signal before PD to make the amplitude lower than A_{SAT}. The proposed method reduces the PAPR at the BS and RS before PD by applying the PAPR reduction utilizing the null space in the MIMO channel. This reduces the signal components whose amplitudes exceed the allowable level for PD and enhances the effectiveness of PD at the BS and RS.

4. Numerical Results

4.1 Simulation Parameters

The performance of the proposed method is evaluated based on computer simulations. Table 1 gives the major simulation parameters. The number of BS transmitter antennas, N_{B}, is set to 100. The number of RS antennas, N_{R}, is 50. The number of UEs, N_{U}, is four. The number of OFDM signal subcarriers, K, is set to 512. The number of frequency blocks, B, is parameterized in the evaluation. The number of FFT/IFFT points, F, is set to 2048, which corresponds to four-times oversampling in the time domain to measure the PAPR levels accurately [42]. To measure the throughput using the Shannon capacity formula, we assume that the symbol constellation at each subcarrier follows an independent complex Gaussian distribution. ZF-based BF is applied. As the channel model, we assume block Rayleigh fading, which is independent between any pairs of transmitter and receiver antennas and between any pair of frequency blocks. For propagation distance D, the distance-dependent path loss is set to $1/D^4$.

The power threshold, T, in the CF process for generating PAPR reduction signals is defined as

$$T = 10 \log_{10}\left(\frac{|A_{\max}|^2}{P_{\mathrm{total}}/FN_{\mathrm{t}}}\right), \ \mathrm{dB} \tag{14}$$

where $A_{\max}$ is the maximum allowable amplitude for clipping the signal, P_{total} is the total transmission power, and N_{t} is the number of transmitter antennas. The power thresholds in the CF process for generating PAPR reduction signals for the BS and RS transmitters are denoted as T_{B} and T_{R}, respectively. The number of iterations of the PAPR reduction process, J, is set to 20 except for the evaluation in Fig. 12, since the PAPR reduction gain of the proposed method is almost unchanged after approximately 20 iterations [39]. In Fig. 12, J is varied to evaluate the performance-complexity tradeoff.

The IBO of the PA is defined as

Table 1 Simulation parameters.

Number of BS antennas, N_{B}	100
Number of RS antennas, N_{R}	50
Number of UEs, N_{U}	4
Number of subcarriers, K	512
Number of FFT/IFFT points, F	2048
Number of iterations, J	20
Constellation of data modulation	i.i.d. complex Gaussian distributed
BF	Zero-forcing
Channel model	Block Rayleigh No fading correlations between any pair of transmitter and receiver antennas and between any pair of frequency blocks
Distance-dependent path loss	$1/D^4$
Throughput calculation	Shannon formula where Bussgang theorem is taken into account

$$\mathrm{IBO} = 10\log_{10}\left(\frac{A_{\mathrm{SAT}}^2}{P_{\mathrm{total}}/FN_{\mathrm{t}}}\right)\mathrm{dB}. \tag{15}$$

The IBOs of the PA at the BS and RS are denoted as $\mathrm{IBO_B}$ and $\mathrm{IBO_R}$, respectively. In the Rapp SSPA model, G_0 is set to 1, A_{SAT} is 0.05, and p is set to 2.0 referring to [43]. The ratio of the received power of the transmission signal from the BS to the noise power at the RS receiver when $\mathrm{IBO_B}$ is set to 0 dB, $\mathrm{SNR_{BR}}$, is defined as

$$\mathrm{SNR_{BR}} = 10\log_{10}\left(\frac{A_{\mathrm{SAT}}^2 FN_{\mathrm{B}}/D_{\mathrm{BR}}^4}{\sum_{b=1}^{B}\sum_{n=1}^{N_{\mathrm{R}}}\left\|\mathbf{z}_{\mathrm{R},b,n}\right\|^2}\right)\mathrm{dB}, \tag{16}$$

where $\mathbf{z}_{\mathrm{R},b,n}$ is the K/B-dimensional noise vector observed at RS receiver antenna n in frequency block b. The ratio of the received power of the transmission signal from the RS to the noise power at the UE receivers when $\mathrm{IBO_R}$ is set to 0 dB, $\mathrm{SNR_{RU}}$, is defined as

$$\mathrm{SNR_{RU}} = 10\log_{10}\left(\frac{A_{\mathrm{SAT}}^2 FN_{\mathrm{R}}/D_{\mathrm{RU}}^4}{\sum_{b=1}^{B}\sum_{n=1}^{N_{\mathrm{U}}}\left\|\mathbf{z}_{\mathrm{U},b,n}\right\|^2}\right)\mathrm{dB}, \tag{17}$$

where $\mathbf{z}_{\mathrm{U},b,n}$ is the K/B-dimensional noise vector observed at UE n in frequency block b.

The PAPR is defined as the ratio of the peak signal power to the average signal power across all the transmitter antennas per OFDM symbol and is expressed as

$$\mathrm{PAPR} = 10\log_{10}\left(\frac{\max_{n,t}|x_n(t)|^2}{P_{\mathrm{total}}/FN_{\mathrm{t}}}\right)\mathrm{dB}, \tag{18}$$

where $x_n(t)$ represents the time domain transmission signal at time t at antenna n.

The data transmission performance is measured by the throughput calculated based on the Shannon formula with the Bussgang theorem to consider appropriately the correlation between the transmitter signal and the PAPR reduction signal, which is explained below in detail. Hereafter, in frequency block b, the K/B-dimensional data stream vector before BF at stream n and the K/B-dimensional received signal vector at UE n are denoted as $\mathbf{s}_{b,k}$ and $\mathbf{y}_{\mathrm{U},b,k}$, respectively. The K/B-dimensional interference and noise signal component vector in the received signal vector at UE n is denoted as $\mathbf{s}_{b,k}$. Defining $\tilde{\mathbf{y}}_{\mathrm{U},b,n}$ and $\mathbf{w}_{b,n}$ as the data signal component vector and the interference and noise signal component vector in $\mathbf{y}_{\mathrm{U},b,k}$, respectively, (4) can be expressed as

$$\begin{aligned}\mathbf{Y}_{\mathrm{U},b} &= \left[\mathbf{y}_{\mathrm{U},b,1}\cdots\mathbf{y}_{\mathrm{U},b,N_{\mathrm{U}}}\right]^T \\ &= A_{\mathrm{B}}\left[\mathbf{s}_{b,1}\cdots\mathbf{s}_{b,N_{\mathrm{U}}}\right]^T + \left[\mathbf{w}_{b,1}\cdots\mathbf{w}_{b,N_{\mathrm{U}}}\right]^T, \\ &= \left[\tilde{\mathbf{y}}_{\mathrm{U},b,1}\cdots\tilde{\mathbf{y}}_{\mathrm{U},b,N_{\mathrm{U}}}\right]^T + \left[\mathbf{w}_{b,1}\cdots\mathbf{w}_{b,N_{\mathrm{U}}}\right]^T\end{aligned} \tag{19}$$

where $\tilde{\mathbf{y}}_{\mathrm{U},b,n} = A_{\mathrm{B}}\mathbf{s}_{b,n}$ holds. As shown in [44], based on the Bussgang theorem, $\mathbf{w}_{b,n}$ can be written as $\mathbf{w}_{b,n} = (\alpha_{b,n}-1)\,\tilde{\mathbf{y}}_{\mathrm{U},b,n} + \mathbf{d}_{b,n}$, where

$$\alpha_{b,n} = \frac{\tilde{\mathbf{y}}_{\mathrm{U},b,n}^H \mathbf{y}_{\mathrm{U},b,n}}{\left\|\tilde{\mathbf{y}}_{\mathrm{U},b,n}\right\|^2} \tag{20}$$

and $\mathbf{d}_{b,n}$ is uncorrelated with $\tilde{\mathbf{y}}_{\mathrm{U},b,n}$. The sum throughput of N_{U} streams (users) is calculated based on the Shannon formula, which corresponds to the upper limit of the transmission rate that can be transmitted without error using ideal channel coding. The sum throughput is expressed as

$$C = \frac{1}{B}\sum_{b=1}^{B}\sum_{n=1}^{N_{\mathrm{U}}}\log_2\left(1+\frac{\left\|\alpha_{b,n}\tilde{\mathbf{y}}_{\mathrm{U},b,n}\right\|^2}{\left\|\mathbf{d}_{b,n}\right\|^2}\right). \tag{21}$$

We mainly evaluate the throughput when the required ACLR is satisfied. Values for T_{B}, T_{R}, $\mathrm{IBO_B}$, and $\mathrm{IBO_R}$ for each method are selected so that the achievable throughput is as high as possible subject to the ACLR level greater than the required value to make a fair comparison. Therefore, some signals input to the PD process may have amplitudes beyond the saturation value, A_{SAT}, that PD cannot address, and the PD cannot perfectly compensate for the nonlinearity of the PA as described in Sects. 2 and 3.2. In ACLR measurements, the assigned channel bandwidth, W, including the guard band is defined as $W = (10/9)K$ in terms of the number of subcarriers, referring to 5G NR specifications with a bandwidth of 5 MHz and subcarrier spacing of 15 kHz [45]. In the simulation assumptions in this paper, W corresponds to approximately 570 subcarriers. The center frequency of the adjacent channel is defined as W away from the center frequency of the assigned channel. The ACLR is defined as

$$\mathrm{ACLR} = 10\log_{10}\frac{P_{\mathrm{C}}}{P_{\mathrm{U}}}\,\mathrm{dB}, \tag{22}$$

where P_C and P_{U} are the average power in the band equivalent to K subcarriers of the assigned channel and the adjacent channel, respectively. In the 5G NR specifications, the ACLR requirement is 45 dB or higher for a BS operating in bands other than bands n46, n96, n102, and n104 [45]. In the following evaluation, the required ACLR is set to 45 dB.

4.2 Simulation Results

Figure 3 shows the complementary cumulative distribution function (CCDF) of the PAPR before the PD process at the BS and RS. In addition to the proposed method, the combination of CF and PD [34] and the PD without PAPR reduction are tested for comparison. Terms T_{B} and T_{R} in the proposed method are set to 2 dB and 3 dB, respectively. Term T_{B} in the CF with PD is set to 2 dB. Terms T_{B} and T_{R} in the proposed method are set to the combination that maximizes the throughput. Terms D_{BR} and D_{RU} are set to 1 and D_{BU} is set to 2. Here, $\mathrm{SNR_{BR}}$ is 10 dB and $\mathrm{SNR_{RU}}$ is 7 dB. Terms $\mathrm{IBO_B}$ and $\mathrm{IBO_R}$ are set to 0 dB. Term B is set to 8. The CF with PD reduces the PAPR at the BS. However, the PAPR at

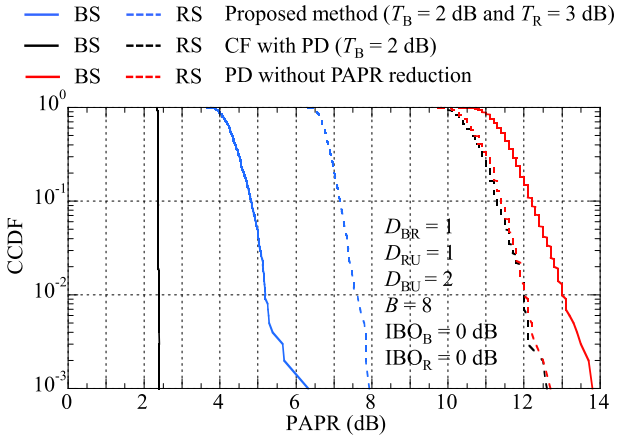


Fig. 3 CCDF of PAPR before PD process at BS and RS.

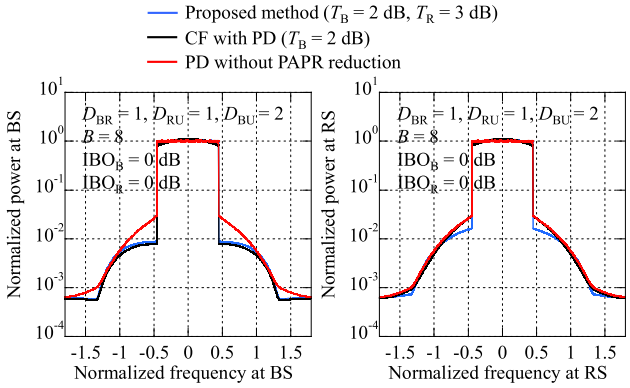


Fig. 4 Transmission signal spectrum.

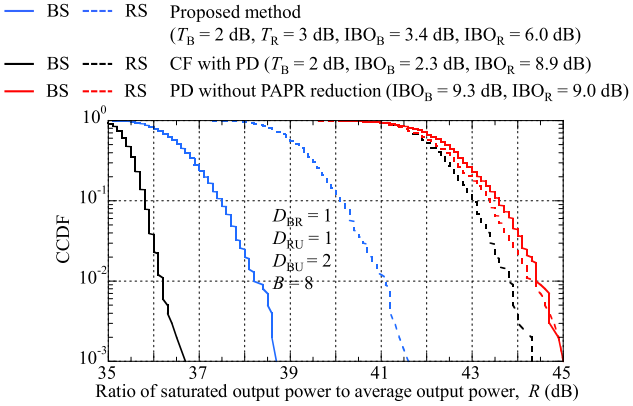


Fig. 5 CCDF of ratio of saturated output power to average output power, R.

the RS in the CF with PD is as high as that in the PD without PAPR reduction. This is because even if only the PAPR at the BS is reduced, the PAPR level of the received signals at the RS is high due to the superposition of the transmission signals from all transmitter antennas of the BS. On the other hand, the proposed method reduces the PAPR levels both at the BS and RS, although the PAPR at the RS is slightly higher than that at the BS due to the increase in PAPR caused by noise added at the RS receiver. The PAPR level at the BS in the proposed method is higher than that in the CF with PD. This is because the projection of the PAPR reduction signal onto the null space of the MIMO channel degrades the PAPR reduction effect at the cost of suppressing the interference to data streams at the UE receivers. These results confirm that the proposed method reduces the PAPR level of signals before the PD process at the BS and RS, although the proposed method does not need any PAPR reduction process at the RS.

Figure 4 shows the transmission signal spectrum at the BS and RS. The signal power normalized by the average power per subcarrier in the band of the assigned channel is represented on the vertical axis. The frequency normalized by the assigned channel bandwidth, W, which includes the guard band is represented on the horizontal axis. In addition to the proposed method, CF with PD and PD without PAPR reduction are tested for comparison. Terms T_B and T_R in the proposed method are set to 2 dB and 3 dB, respectively. Term T_B in the CF with PD is set to 2 dB. Terms D_{BR} and D_{RU} are set to 1 and D_{BU} is set to 2. The SNR_{BR} is 10 dB and SNR_{RU} is 7 dB. Terms IBO_B and IBO_R are set to 0 dB. Term B is set to 8. The proposed method has less out-of-band spectral leakage at the BS and RS than that for the PD without PAPR reduction. This is because the proposed method has lower PAPR levels at the BS and RS and thereby increases the effectiveness of the PD at the BS and RS. Since the proposed method has a higher PAPR at the RS than at the BS, the degree of enhancement of the PD effect at the RS is lower than that at the BS, resulting in a greater out-of-band spectral leakage at the RS than at the BS. The proposed method has slightly greater out-of-band spectral leakage at the BS than that for the CF with PD. This is because the proposed method has a higher PAPR level before the PD at the BS than that for the CF with PD. However, the CF with PD has the same degree of out-of-band spectral leakage at the RS as that for the PD without PAPR reduction. This is because the CF with PD has as high PAPR level before the PD at the RS as that for the PD without PAPR reduction as shown in Fig. 3. Therefore, the CF with PD greatly weakens the effectiveness of PD at the RS. These results confirm that the proposed method suppresses out-of-band spectral leakage using PAPR reduction on both the BS and RS ends thanks to the enhanced PD effect.

When out-of-band spectral leakage is reduced, the IBO to satisfy the required ACLR can be reduced and the transmission power can be increased as a result. Figure 5 shows the CCDF of the ratio, R, of the saturated output power $(G_0 A_{SAT})^2$ to average output power. A lower R indicates a higher output power of the PAs. In addition to the proposed method, the CF with PD and the PD without PAPR reduction are tested for comparison. Terms T_B, T_R, IBO_B, and IBO_R are set to the combination that maximizes throughput with an ACLR of 45 dB or higher for each method. Terms D_{BR} and D_{RU} are set to 1 and D_{BU} is set to 2. The SNR_{BR} is 10 dB and SNR_{RU} is 7 dB. Term B is set to 8. The proposed method reduces R at both the BS and RS compared to

PD without PAPR reduction. This is because the proposed method improves the out-of-band spectral leakage at both the BS and RS, thereby reducing the IBO_B and IBO_R to satisfy the required ACLR. In the proposed method, R at the RS is higher than that at the BS. This is because the proposed method improves the out-of-band spectral leakage at the BS more significantly than at the RS and reduces IBO_B more than IBO_R. The CF with PD has a lower R at the BS than that for the proposed method. However, the level of R at the RS in the CF with PD is high which is close to that in the PD without PAPR reduction. These results confirm that the proposed method increases the output power of the PAs at both the BS and RS when the ACLR is restricted.

Figures 6 and 7 show the average throughput, IBO_B, and IBO_R as a function of the number of frequency blocks, B, for the case when PD is applied or not, respectively. In Fig. 7, the PAPR reduction method used in the proposed method (denoted as "PAPR reduction using null space of MIMO channel" in Fig. 7), which is described in Sect. 3.1, and that used in the CF with PD (denoted as "CF" in Fig. 7) are tested. Terms D_{BR} and D_{RU} are set to 1 and D_{BU} is set to 2. The SNR_{BR} is 10 dB and SNR_{RU} is 7 dB. Comparing Figs. 6 and 7, we confirm that the PD process decreases the IBO and increases the throughput for both cases with and without PAPR reduction. This is due to the suppression of out-of-band spectral leakage using PD, which increases the ACLR and reduces the IBO satisfying the required ACLR. The increase in throughput through the use of PD is greater when PAPR is reduced than when PAPR is not reduced. This is because the reduction in PAPR reduces the signal components with input amplitudes above the saturation level that the PD cannot address. The proposed method exhibits the largest increase in throughput by using PD. This is because the proposed method enhances the PD effect at both the BS and RS, while the CF with PD enhances the PD effect only at the BS. The PD without PAPR reduction cannot achieve a sufficient PD effect at both the BS and RS.

In both Figs. 6 and 7, as B increases, IBO_B and IBO_R decrease and the throughput increases. The reason for this characteristic is explained in [28]. Reference [28] shows that the reduction in the cross correlation of the transmission signals among transmitter antennas contributes to the enhancement of the effect of the PAPR reduction method using the null space in the MIMO channel, since more PAPR reduction signal components appear in the null space of the MIMO channel. Furthermore, [28] shows that as the frequency selectivity of the channel is increased, which corresponds to the increased B, the cross correlation of the transmission signals among transmitter antennas decreases. Therefore, the throughput increases as B increases. The increase in throughput as B increases is greater in Fig. 6 than in Fig. 7. This is because the effectiveness of PD is enhanced by the lower PAPR level due to the improved PAPR reduction effect.

Figures 8 and 9 show the average throughput, IBO_B, and IBO_R as a function of the distance between the BS and UEs, D_{BU}, for the case when PD is applied or not, respectively. Terms D_{BR} and D_{RU} are set to 1. Here, SNR_{BR} is 10 dB and SNR_{RU} is 7 dB. Term B is set to 8. Overall, the throughput levels of all methods are increased as D_{BU} decreases thanks to the increased direct link gain. Comparing Figs. 8 and 9, the PD process increases the throughput. The proposed method always achieves higher throughput than the other methods, regardless of D_{BU}. This is because the proposed method reduces the PAPR at both the BS and RS while suppressing interference from the PAPR reduction signal to the data stream. In particular, the difference in throughput between the proposed method and the other methods in Fig. 8 is greater than that in Fig. 9. This is because the proposed method enhances the effectiveness of PD at both the BS and RS. These results show that PD is an important technology for coverage enhancement and that the proposed method is effective in enhancing coverage because distance D_{BU} that satisfies the required throughput in the proposed method is greater than that in the CF with PD.

To confirm the effectiveness of the proposed method in enhancing coverage, the throughput is measured when the SNR is varied. Terms D_{BR}, D_{RU}, and D_{BU} are fixed to $\sqrt[4]{2}$, 1, and $1 + \sqrt[4]{2}$, respectively, so that SNR_{BR} and SNR_{RU} have the same value. In the following, SNR_0 is defined as $SNR_0 = SNR_{BR} = SNR_{RU}$. Figures 10 and 11 show the av-

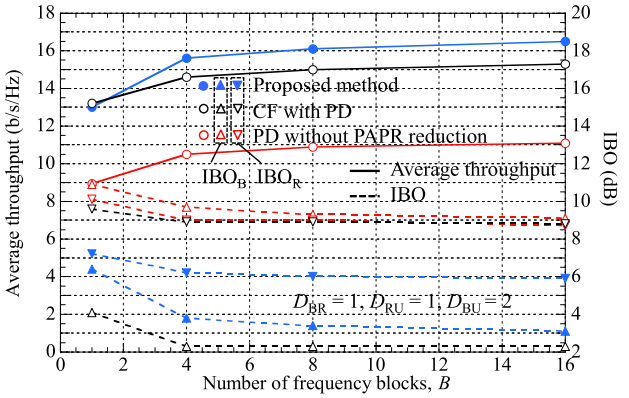


Fig. 6 Average throughput and IBO as a function of number of frequency blocks, B, when PD is applied.

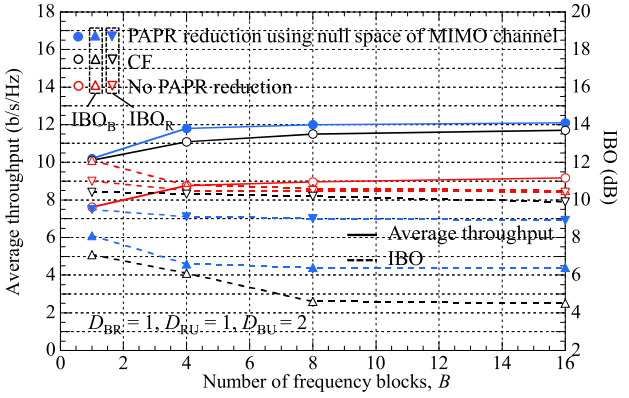


Fig. 7 Average throughput and IBO as a function of number of frequency blocks, B, when PD is not applied.

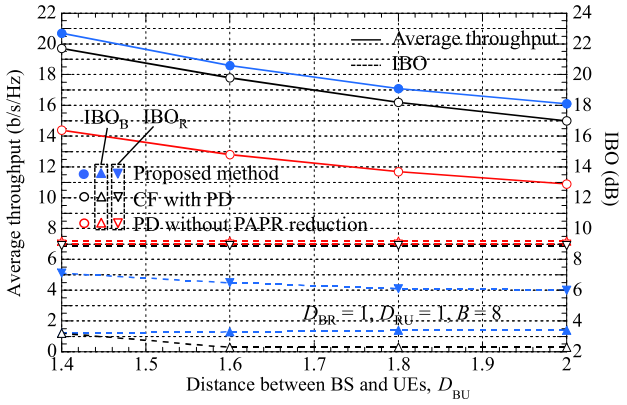


Fig. 8 Average throughput and IBO as a function of distance between BS and UEs, D_{BU}, when PD is applied.

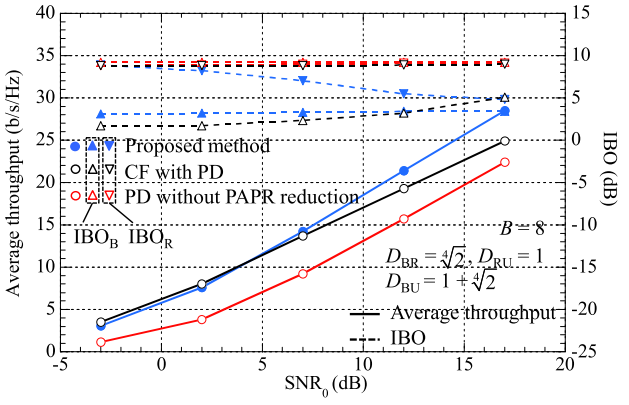


Fig. 10 Average throughput and IBO as a function of SNR_0 when PD is applied.

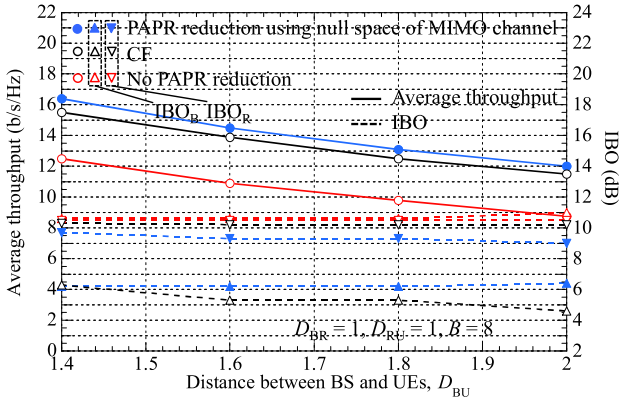


Fig. 9 Average throughput and IBO as a function of distance between BS and UEs, D_{BU}, when PD is not applied.

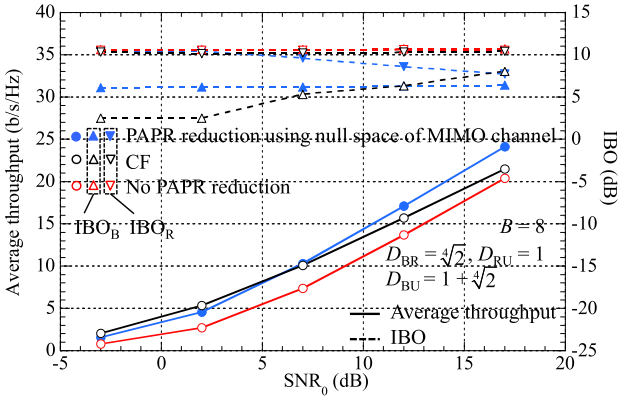


Fig. 11 Average throughput and IBO as a function of SNR_0 when PD is not applied.

erage throughput, $\mathrm{IBO_B}$, and $\mathrm{IBO_R}$ as a function of SNR_0, for the case when PD is applied or not, respectively. Term B is set to 8. Overall, the throughput levels of all methods are increased as SNR_0 increases. Comparing Figs. 10 and 11, we see that the PD process increases the throughput. When the SNR_0 level is high, the proposed method achieves a higher throughput than that for the CF with PD. On the other hand, when the SNR_0 level is very low, the throughput of the proposed method is slightly lower than that for the CF with PD. This is because the proposed method cannot address the increase in PAPR due to noise added at the RS receiver and the PAPR reduction effect of the CF is higher than that for the PAPR reduction method using the null space in the MIMO channel. In Fig. 10, the proposed method and the CF with PD have the same throughput level when SNR_0 is approximately 4 dB. On the other hand, in Fig. 11, the PAPR method using the null space in the MIMO channel and the CF method have the same throughput level when SNR_0 is approximately 6 dB. This indicates that the proposed method achieves lower $\mathrm{IBO_B}$ and $\mathrm{IBO_R}$ levels, and higher throughput by enhancing the effectiveness of the PD at both the BS and RS. Figure 10 shows that as the target throughput level is set higher, the proposed method tends to lower the required SNR_0 compared to the CF with PD and the PD without PAPR reduction. Comparing this to Fig. 11, this trend is more pronounced in Fig. 10. This is because the higher the effective SNR, the smaller the increase in PAPR due to noise added at the RS receiver, and the PAPR reduction method utilizing the null space in the MIMO channel which is used in the proposed method can sufficiently reduce the PAPR at the RS. This allows the proposed method to enhance the effectiveness of PD at the RS. In conclusion, the proposed method provides wider coverage than that for the conventional methods especially when the required throughput is high.

The required calculation cost per iteration depends on the PAPR reduction method. In the following, we compare the throughput of each method for the same computational complexity. The required number of real multiplications is used as a measure of the computational complexity. Table 2 gives the number of real multiplications per iteration for each PAPR reduction method.

Based on Table 2, Fig. 12 shows the average throughput, $\mathrm{IBO_B}$, and $\mathrm{IBO_R}$ as a function of real multiplications for the case when PD is applied. Terms D_{BR} and D_{RU} are set to 1 and D_{BU} is set to 2. Here, $\mathrm{SNR_{BR}}$ is 10 dB and $\mathrm{SNR_{RU}}$ is 7 dB. Term B is set to 8. The PAPR reduction using the null space of the MIMO channel requires a larger number of

Table 2 Number of real multiplications for each PAPR reduction method.

PAPR Reduction Method	Process	Number of Real Multiplications
PAPR reduction using null space of MIMO channel	Generation of Moore-Penrose inverse of channel matrix between BS and RS	$4N_{\mathrm{R}}^2(N_{\mathrm{R}}+2N_{\mathrm{B}})B$
	Calculation of null space matrix and Moore-Penrose inverse	$4N_{\mathrm{B}}^2(N_{\mathrm{B}}+N_{\mathrm{R}}-N_{\mathrm{U}})$
	Prediction of received signal at RS	$4JN_{\mathrm{B}}FN_{\mathrm{R}}$
	IFFT	$4J(N_{\mathrm{B}}+N_{\mathrm{R}})F\log_2 F$
	Power measurement of transmission signal	$2J(N_{\mathrm{B}}+N_{\mathrm{R}})F$
	Amplitude clipping	$3J(N_{\mathrm{B}}+N_{\mathrm{R}})F$
	FFT	$4J(N_{\mathrm{B}}+N_{\mathrm{R}})F\log_2 F$
	Projection onto null space	$4JN_{\mathrm{B}}F(N_{\mathrm{B}}+N_{\mathrm{R}})$
CF	IFFT	$4JN_{\mathrm{B}}F\log_2 F$
	Power measurement of transmission signal	$2JN_{\mathrm{B}}F$
	Amplitude clipping	$3JN_{\mathrm{B}}F$
	FFT	$4JN_{\mathrm{B}}F\log_2 F$

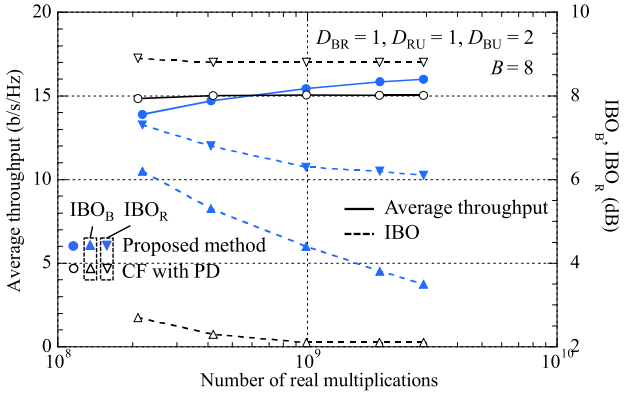


Fig. 12 Average throughput and IBO as a function of number of real multiplications when PD is applied.

real multiplications per iteration than that for CF as shown in Table 2. However, Fig. 12 shows that when the BS can tolerate approximately three or more iterations of PAPR reduction using the null space of the MIMO channel, which is equivalent to approximately 31 or more iterations of CF, the proposed method achieves higher throughput than that for the CF with PD.

5. Conclusion

Our previously reported PAPR reduction method achieves PAPR reduction both at the BS and RS transmitters utilizing the null space in the integrated MIMO channel of the entire system by only using the signal processing at the BS. In this paper, we propose combined use of the previously reported PAPR reduction method with PD in downlink AF-type MIMO-OFDM relay transmission for further performance enhancement. In the proposed method, the BS alternately and repeatedly generates the signal to reduce the PAPR for the BS transmitter and the signal to reduce the PAPR for the RS transmitter, and projects them onto the null space in the integrated MIMO channel of the entire system to suppress the interference to data streams. After the PAPR reduction process, the transmission signals are processed using PD and then input to the PAs of the BS. In addition, the received signals at the RS are processed using PD and input to the PAs of the RS. The proposed method reduces the PAPR at both the BS and RS without any PAPR reduction signal processing at the RS, and this enhances the effectiveness of PD both at the BS and RS. Thus, the proposed method addresses all the drawbacks of the conventional method for multi-antenna relay transmission, that is the interference to the data stream due to the PAPR reduction signal at the UE receiver, high PAPR at the RS, and degradation of the effectiveness of PD at the RS due to a high PAPR. Computer simulation results showed that the proposed method achieves higher throughput and coverage enhancement compared to those for the conventional methods.

References

[1] T.L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," IEEE Trans. Wireless Commun., vol.9, no.11, pp.3590–3600, Nov. 2010.
[2] H. Papadopoulos, C. Wang, O. Bursalioglu, X. Hou, and Y. Kishiyama, "Massive MIMO technologies and challenges towards 5G," IEICE Trans. Commun., vol.E99-B, no.3, pp.602–621, March 2016.
[3] E. Dahlman, S. Parkvall, and J. Sköld, 5G NR: The Next Generation Wireless Access Technology, Academic Press, 2018.
[4] NTT DOCOMO, "White paper: 5G evolution and 6G," Jan. 2023.
[5] R.U. Nabar, H. Bolcskei, and F.W. Kneubuhler, "Fading relay channels: Performance limits and space-time signal design," IEEE J. Sel. Areas Commun., vol.22, no.6, pp.1099–1109, Aug. 2004.
[6] H. Bolcskei, R.U. Nabar, O. Oyman, and A.J. Paulraj, "Capacity scaling laws in MIMO relay networks," IEEE Trans. Wireless Commun., vol.5, no.6, pp.1433–1444, June 2006.
[7] K. Tateishi and K. Higuchi, "Adaptive amplify-and-forward relaying for cellular downlink" IEICE Trans. Commun., vol.E96-B, no.7, pp.1968–1975, July 2013.
[8] G. Tong, Q. Wang, and G. Wang, "A performance controllable PA linearization scheme of joint PAPR reduction and predistortion," 2012 IEEE ICCT2012, pp.1177–1181, Chengdu, Nov. 2012.
[9] S.H. Han and J.H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," IEEE Wireless Commun., vol.12, no.2, pp.56–65, April 2005.
[10] X. Li and L.J. Cimini, Jr., "Effect of clipping and filtering on the performance of OFDM," IEEE Commun. Lett., vol.2, no.5, pp.131–133, May 1998.
[11] J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," Elect. Lett., vol.38, no.8, pp.246–247, Feb. 2002.
[12] B.S. Krongold and D.L. Jones, "PAR reduction in OFDM via active constellation extension," IEEE Trans. Broadcast., vol.49, no.3, pp.258–268, Sept. 2003.
[13] A. Aggarwal and T.H. Meng, "Minimizing the peak-to-average power ratio of OFDM signals using convex optimization," IEEE Trans. Signal Process., vol.54, no.8, pp.3099–3110, Aug. 2006.
[14] J. Tellado and J.M. Cioffi, "Efficient algorithms for reducing PAR in multicarrier systems," Proc. IEEE Int. Symp. Inf. Theory, p.191, Cambridge, MA, Aug. 1998.
[15] H. Lee, D.N. Liu, W. Zhu, and M.P. Fitz, "Peak power reduction using a unitary rotation in multiple transmit antennas," Proc. IEEE ICC2005, pp.2407–2411, Seoul, Korea, May 2005.
[16] G.R. Woo and D.L. Jones, "Peak power reduction in MIMO OFDM via active channel extension," Proc. IEEE ICC2005, pp.2636–2639, Seoul, Korea, May 2005.
[17] S. Suyama, H. Adachi, H. Suzuki, and K. Fukawa, "PAPR reduction methods for eigenmode MIMO-OFDM transmission," Proc. IEEE

VTC2009-Spring, Barcelona, Spain, April 2009.
[18] H. Ando and K. Higuchi, "Comparison of PAPR reduction methods for OFDM signal with channel coding," Proc. IEEE APWCS2009, Seoul, Korea, Aug. 2009.
[19] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek, "A low-complex peak-to-average power reduction scheme for OFDM based massive MIMO systems," Proc. ISCCSP2014, Athens, Greece, May 2014.
[20] T. Kageyama, O. Muta, and H. Gacanin, "Enhanced peak cancellation with simplified in-band distortion compensation for massive MIMO-OFDM," IEEE Access, vol.8, pp.73420–73431, July 2020.
[21] M. Iwasaki and K. Higuchi, "Clipping and filtering-based PAPR reduction method for precoded OFDM-MIMO signals," Proc. IEEE VTC2010-Spring, Taipei, Taiwan, May 2010.
[22] Y. Sato, M. Iwasaki, S. Inoue, and K. Higuchi, "Clipping and filtering-based adaptive PAPR reduction method for precoded OFDM-MIMO signals," IEICE Trans. Commun., vol.E96-B, no.9, pp.2270–2280, Sept. 2013.
[23] S. Inoue, T. Kawamura, and K. Higuchi, "Throughput/ACLR performance of CF-based adaptive PAPR reduction method for eigenmode MIMO-OFDM signals with AMC," IEICE Trans. Commun., vol.E96-B, no.9, pp.2293–2300, Sept. 2013.
[24] R. Kimura, Y. Tajika, and K. Higuchi, "CF-based adaptive PAPR reduction method for block diagonalization-based multiuser MIMO-OFDM signals," Proc. IEEE VTC2011-Spring, Budapest, Hungary, May 2011.
[25] Y. Matsumoto, K. Tateishi, and K. Higuchi, "Performance evaluations on adaptive PAPR reduction method using null space in MIMO channel for eigenmode massive MIMO-OFDM signals," Proc. APCC2017, Perth, Australia, Dec. 2017.
[26] T. Suzuki, M. Suzuki, Y. Kishiyama, and K. Higuchi, "Complexity-reduced adaptive PAPR reduction method using null space in MIMO channel for MIMO-OFDM signals," IEICE Trans. Commun., vol.E103-B, no.9, pp.1019–1029, Sept. 2020.
[27] T. Suzuki, M. Suzuki, and K. Higuchi, "Parallel peak cancellation signal-based PAPR reduction method using null space in MIMO channel for MIMO-OFDM transmission," IEICE Trans. Commun., vol.E104-B, no.5, pp.539–549, May 2021.
[28] L. Yamaguchi, N. Nonaka, and K. Higuchi, "PC-signal-based PAPR reduction using null space in MIMO channel for MIMO-OFDM signals in frequency-selective fading channel," Proc. IEEE VTC2020-Fall, Virtual Conference, Nov.-Dec. 2020.
[29] S. Shin, N. Nonaka, and K. Higuchi, "User group selection method in multiuser MIMO-OFDM transmission with adaptive PAPR reduction using null space in MIMO channel," Proc. IEEE VTC2020-Fall, Virtual Conference, Nov.-Dec. 2020.
[30] J. Saito, N. Nonaka, and K. Higuchi, "PAPR reduction using null space in MIMO channel considering signal power difference among transmitter antennas," Proc. IEEE WCNC2023, Glasgow, Scotland, UK, March 2023.
[31] R. Zayani, H. Shaiek, and D. Roviras, "PAPR-aware massive MIMO-OFDM downlink," IEEE Access, vol.7, pp.25474–25484, Feb. 2019.
[32] L. Hua, Y. Wang, Z. Lian, Y. Su, and Z. Xie, "Low-complexity PAPR-aware precoding for massive MIMO-OFDM downlink systems," IEEE Wireless Commun. Lett., vol.11, no.7, pp.1339–1343, July 2022.
[33] M. Eddaghel and J.A. Chambers, "PAPR reduction in distributed amplify-and-forward type closed loop extended orthogonal space frequency block coding with one-bit group feedback for cooperative communications," European Wireless 2012, pp.1–6, Poznan, Poland, 2012.
[34] M.D.N. Habibah, G.S. Palupi, A.A. Puspitasari, U.A. Nadhiroh, M. Ridwan, and Y. Moegiharto, "Performance of a joint PAPR reduction clipping and filtering (CF) scheme and predistortion techniques in amplify and forward (AF) relaying system with relay selection strategy," IES2021, pp.120–125, Surabaya, Indonesia, Sept. 2021.
[35] S. Yang, W. Yang, Y. Cai, and W. Li, "An energy efficient PTS scheme for PAPR reduction in OFDM relay systems," Proc. ChinaCom2015, pp.858–863, Shanghai, China, Aug. 2015.
[36] M.N. Saniar, N.P. Anggraini, Arifin, and Y. Moegiharto, "Evaluation of the PTS PAPR reduction technique with the Hammerstein-Wiener predistortion model in amplify-and-forward (AF), decode-and-forward (DF) relaying systems over asymmetric channels," IES2021, pp.87–91, Surabaya, Indonesia, Sept. 2021.
[37] Y. Sekiguchi, N. Nonaka, and K. Higuchi, "PAPR reduction using null space in MIMO channel for MIMO-OFDM signals in multiple-antenna AF relay transmission," Proc. IEEE VTC2021-Fall, Online, Sept.-Oct. 2021.
[38] Y. Sekiguchi, N. Nonaka, and K. Higuchi, "PAPR reduction of OFDM signals using null space in MIMO channel for MIMO amplify-and-forward relay transmission," IEICE Trans. Commun., vol.E105-B, no.9, pp.1078–1086, Sept. 2022.
[39] A. Kakehashi, N. Nonaka, and K. Higuchi, "PAPR reduction using null space in MIMO channel based on signal processing at base station for downlink AF-based relaying MIMO-OFDM signals," Proc. IEEE VTC2022-Fall, Online, Sep.-Oct. 2022.
[40] A. Kakehashi, T. Hara, and K. Higuchi, "Base station-driven PAPR reduction method utilizing null space for MIMO-OFDM systems with amplify-and-forward relaying," IEEE Access, vol.12, pp.24714–24724, Feb. 2024.
[41] C. Dudak, A.T. Koc, and S. Koc, "Solid state power amplifier (SSPA) non-linearity effects on quadri-phase shift keying modulation," Proc. EuWiT2004, Amsterdam, Netherlands, Oct. 2004.
[42] M. Sharif, M. Gharavi-Alkhansari, and B.H. Khalaj, "On the peak-to-average power of OFDM signals based on oversampling," IEEE Trans. Commun., vol.51, no.1, pp.72–78, Jan. 2003.
[43] E. Costa and S. Pupolin, "M-QAM-OFDM system performance in the presence of a non-linear amplifier and phase noise," IEEE Trans. Commun., vol.50, no.3, pp.462–472, March 2002.
[44] H. Ochiai and H. Imai, "Performance analysis of deliberately clipped OFDM signals," IEEE Trans. Commun., vol.50, no.1, pp.89–101, Jan. 2002.
[45] 3GPP TS38.104, NR Base Station (BS) radio transmission and reception (Release 17).

Asuka Kakehashi received the B.E. and M.E. degrees from Tokyo University of Science, Noda, Japan in 2022 and 2024, respectively. In 2024, he joined NEC Corporation. His research interests include wireless communications.

Kenichi Higuchi received the B.E. degree from Waseda University, Tokyo, Japan, in 1994, and received the Dr.Eng. degree from Tohoku University, Sendai, Japan in 2002. In 1994, he joined NTT Mobile Communications Network, Inc. (now, NTT DOCOMO, INC.). While with NTT DOCOMO, he was engaged in the research and standardization of wireless access technologies for wideband DS-CDMA mobile radio, HSPA, LTE, and broadband wireless packet access technologies for systems beyond IMT-2000. In 2007, he joined the faculty of the Tokyo University of Science and currently holds the position of Professor. His current research interests are in the areas of wireless technologies and mobile communication systems, including advanced multiple access such as non-orthogonal multiple access (NOMA), radio resource allocation, inter-cell interference coordination, multiple-antenna transmission techniques, signal processing such as interference cancellation and turbo equalization, and issues related to heterogeneous networks using small cells. He was a co-recipient of the Best Paper Award of the International Symposium on Wireless Personal Multimedia Communications in 2004 and 2007, the Best Paper Award from the IEICE in 2021, a recipient of the Young Researcher's Award from the IEICE in 2003, the 5th YRP Award in 2007, the Prime Minister Invention Prize in 2010, and the Invention Prize of Commissioner of the Japan Patent Office in 2015. He is a senior member of the IEEE.

IEICE TRANS. COMMUN., VOL.E108–B, NO.1 JANUARY 2025

PAPER

SAR Image Generation of 3D Target with Consideration of Complex RCS*

Xian YU† *and* **Yubing HAN**†a), *Nonmembers*

SUMMARY Synthetic aperture radar (SAR) image generation is crucial to SAR image interpretation when sufficient image samples are unavailable. Against this background, a method for SAR image generation of three-dimensional (3D) target is proposed in this paper. Specifically, this method contains three steps. Firstly, according to the system parameters, the echo signal in the two-dimensional (2D) time domain is generated, based on which 2D Fast Fourier Transform (2DFFT) is performed. Secondly, the hybrid moments (MoM)-large element physical optics (LEPO) method is used to calculate the scattering characteristics with the certain frequency points and incident angles according to the system parameters. Finally, range Doppler algorithm (RDA) is adopted to process the signal in the 2D-frequency domain with radar cross section (RCS) exported from electromagnetic calculations. These procedures combine RCS computations by FKEO solver and RDA to simulate raw echo signal and then generate SAR image samples for different squint angles and targets with reduced computational load, laying foundations for transmit waveform design, SAR image interpretation and other SAR related work.

key words: *SAR image generation, RCS, 2DFFT, electromagnetic calculations, hybrid MoM-LEPO method*

1. Introduction

Synthetic aperture radar (SAR) refers to a certain radar system which can provide high-resolution two dimensional (2D) images in all-weather and all-time conditions [1]–[5]. Accordingly, SAR systems have been widely applied in several scenarios, e.g., resource exploration, SAR-communication integration, and military surveillance [3], [6]. Among these SAR related research, SAR image generation based on raw echo simulation is of ever-increasing importance. This is due to that on the one hand, SAR raw echo simulation taking target radar cross section (RCS) features into consideration is economic in verifying the effectiveness of waveform design and imaging algorithm; on the other hand, SAR image generation is crucial to establish a target SAR image database, which is a basic premise of SAR image interpretation including target detection and identification.

In recent literatures, several SAR image generation methods have been proposed. For instance, in [7], classic time-domain (TD) method was employed to simulate SAR raw signals, which is proved to be valid and precise but time-consuming. Additionally, with the rapid development of neural networks, the adversarial auto-encoder network aiming at few shot learning was introduced into SAR image generation in [8], [9], while this kind of network can only generate images according to the existing SAR samples but fail to generate novel SAR images. Xia, et al. studied raw echo simulation of manoeuvring targets for missile-borne SAR based on one-dimensional Fast Fourier transform (1DFFT) algorithm using the physical optics (PO)/geometrical optics (GO) and incremental length diffraction coefficients (ILDC) based RCS calculation method [10]. Similarly, [11] combines 1DFFT with time-domain shooting and bouncing ray (TDSBR) algorithm to simulate SAR raw echo signal. Overall, this 1DFFT method is computationally efficient compared with TD method, but more time-consuming than 2D frequency domain method proposed in [12]. Specifically, [12] focuses on the side-looking imaging in 2D frequency domain without consideration of target RCS as well as small and high squint angles. Furthermore, among state-of-the-art electromagnetic techniques, applying Stochastic Galerkin Method (SGM) to integral equations combined with the method of moments (MoM) has high accuracy, but requires solving a large deterministic problem [13]. Considering the high complexity and time-consuming computation of electrically large targets, large element physical optics (LEPO) method is exploited to calculate complex RCS owing to the characteristics of low complexity, faster calculation and applicability [14]–[16]. [17] elaborated in detail on the hybrid theory of MoM and PO, and successfully applied it in complex electromagnetic scattering calculations. Furthermore, considering the high complexity of electromagnetic calculations [18], FEKO as the first commercial tool for electromagnetic field analysis of 3D structures [19] is introduced to calculate RCS features. Nevertheless, the combination of RCS calculations and SAR imaging in 2D frequency domain is still missing.

To overcome the above issues, a framework of SAR image generation of 3D target with consideration of complex RCS is proposed in this paper, laying foundations for waveform design and SAR image interpretation. The contributions are listed as follows:

- The proposed method is divided into three steps, which are: (1) Simulation for raw echo in the 2D frequency domain; (2) MoM-LEPO-based RCS calculation; (3) SAR image acquisition with the matched filters on the target echo composed by step (1) and (2).
- To obtain reasonable SAR images of 3D targets, RCS is calculated via hybrid MoM-LEPO method with the parameters set according to the synthetic angle in SAR

Manuscript received February 5, 2024.
Manuscript publicized June 28, 2024.

†School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

*This work is supported by the National Natural Science Foundation of China under Grant 62371236.

a) E-mail: hanyb@njust.edu.cn

DOI: 10.23919/transcom.2024EBP3027

imaging system.
- Target echo with complex RCS is simulated in the 2D frequency domain, which reduces the computational load.

The remainder of the paper is organized as follows. In Sect. 2, the system model is presented, and procedures of RCS calculations and echo signal processing are also introduced. The overall flowchart is given. In Sect. 3, simulations are given to verify the effectiveness of the proposed SAR image generation. Section 4 draws the conclusion.

2. Analysis of the Proposed SAR Image Generation

2.1 Raw Echo Signal Simulation

The geometry of the airborne strip-map SAR system considered to work in the squint-looking mode is presented in Fig. 1.

As shown in Fig. 1, we assume that the radar flies at the constant height of H with the speed of v along the positive Y-direction at the depression angle θ and squint angle φ. The depression angle θ is measured from the Z-axis, and the squint angle φ is measured from the XOZ plane. We also assume that the target P, located at $(x_p, y_p, 0)$, is the radiation centre with the initial range R_0 and the instantaneous range between radar and the target is denoted as $R(t_a)$. The transmitted chirp signal is written as

$$s(t_r, t_a) = rect(t_r/T)\cdot w_a(t_a)\cdot \exp\left(j\pi\mu {t_r}^2\right)\cdot \exp(j2\pi f_c t_r), \tag{1}$$

where t_r is the fast time, T is the pulse repetition time and $rect(t_r/T)$ denotes the rectangle function, written as

$$rect(t_r/T) = \begin{cases} 1 & |t_r/T| \le 0.5, \\ 0 & \text{else}, \end{cases} \tag{2}$$

where $|\cdot|$ means the modulus operation. Also, t_a is the slow time, $w_a(t_a)$ is the azimuth time envelope, μ is the chirp rate, f_c is the carrier frequency.

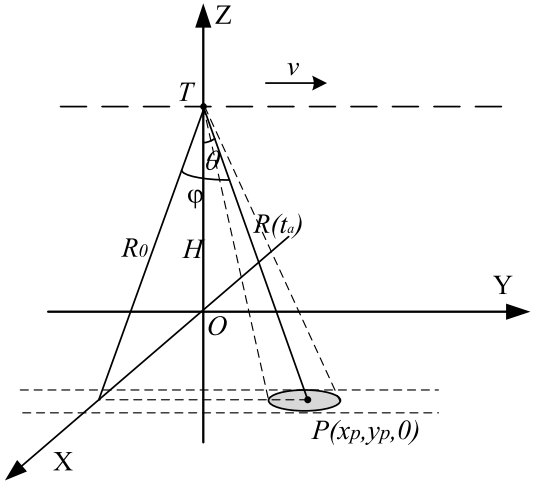


Fig. 1 Geometric configuration of squint-looking airborne-SAR.

After demodulation to baseband, the echo reflected from the scatters can be denoted as

$$s_R(t_r, t_a, \theta, \varphi) = \sigma(t_r, \theta, \varphi) \otimes s\left(t_r - \frac{2R(t_a)}{c}, t_a\right), \tag{3}$$

where $\sigma(t_r, \theta, \varphi)$ represents the reflectivity function of t_r, the depression angle θ and squint angle φ, $\otimes$ denotes convolution operation, c is the speed of light, and $s\left(t_r - \frac{2R(t_a)}{c}, t_a\right)$ is expressed as

$$\begin{aligned} s\left(t_r - \frac{2R(t_a)}{c}, t_a\right) =& rect\left(\frac{t_r - 2R(t_a)/c}{T}\right)\cdot w_a(t_a)\cdot \\ & \exp\left(j\pi\mu\left(t_r - \frac{2R(t_a)}{c}\right)^2\right)\cdot \exp\left(-j4\pi f_c \frac{R(t_a)}{c}\right) \end{aligned} \tag{4}$$

and $R(t_a)$ can be given as $R(t_a) = \sqrt{{R_0}^2 + v^2 {t_a}^2}$. Due to the heavy computational load of convolution in the time domain, 2DFFT is performed on (3) to convert time-domain convolution into frequency-domain multiplication so as to obtain the echo signal in the 2D frequency domain. The process of solving the 2D frequency domain expression of $s\left(t_r - \frac{2R(t_a)}{c}, t_a\right)$ is as follows.

Applying the azimuth Fourier transform into (3) and using the principle of stationary phase (POSP) yields

$$\begin{aligned} S_1(f_r, t_a, \theta, \varphi) =& \sigma(f_r, \theta, \varphi)\cdot W_r(f_r)\cdot w_a(t_a)\cdot \\ & \exp\left(-j\frac{4\pi(f_c+f_r)R(t_a)}{c}\right)\cdot \exp\left(-j\frac{\pi {f_r}^2}{\mu}\right), \end{aligned} \tag{5}$$

where $W_r(f_r)$ represents the range frequency envelope. Furthermore, it can be seen in (5) that the complex RCS $\sigma(f_r, \theta, \varphi)$ correspond to the frequency response of the scatters over frequency series $f_c + f_r$. And then, transforming (5) into the 2-D frequency domain, it can be denoted as

$$\begin{aligned} S_2(f_r, f_a, \theta, \varphi) =& A(f_r, \theta, \varphi)\cdot W_r(f_r)\cdot W_a(f_a)\cdot \\ & \exp(j\Theta(f_r, f_a))\cdot \exp\left(-j\frac{\pi {f_r}^2}{\mu}\right), \end{aligned} \tag{6}$$

where $A(f_r, \theta, \varphi)$ denotes the frequency response of the scatters in both range and azimuth directions, $W_a(f_a)$ is the azimuth frequency envelope, and $\Theta(f_r, f_a)$ is written as

$$\Theta(f_r, f_a) = -\frac{4\pi f_c R_0}{c}\sqrt{D(f_a, v)^2 + \frac{2f_r}{f_c} + \frac{{f_r}^2}{{f_c}^2}}, \tag{7}$$

where

$$D(f_a, v) = \sqrt{1 - \frac{c^2 {f_a}^2}{4v^2 {f_c}^2}}. \tag{8}$$

For further analysis, (7) can be expanded to Taylor's series at $f_r = 0$

$$\Theta(f_r, f_a) = -\frac{4\pi f_c R_0}{c} \left(D(f_a, v) + \frac{f_r}{f_c D(f_a, v)} - \frac{{f_r}^2}{2{f_c}^2 D(f_a, v)^3} \frac{c^2 {f_a}^2}{4v^2 {f_c}^2} \right). \quad (9)$$

It follows then that the items in (6) apart from $A(f_r, \theta, \varphi)$ can be simulated according to the system parameters. Therefore, RCS calculations are crucial in echo simulation of target in the 2D frequency domain. Based on complex RCS, it is noted that LFM can be replaced by other waveforms by changing the weights of different frequency points through FFT on the waveform. In this paper, we take LFM as an example and design the corresponding matched filters.

2.2 MoM-LEPO-Based RCS Calculations

Turntable imaging is a powerful tool for the calculations of scattering properties of targets [20], [21]. The scattering characteristics are simulated over the depression angle θ and rotation angle α centring on the squint angle φ. The top view is given in Fig. 2. Specifically, the down-range is defined as the axis parallel to the direction of incidence, with a frequency series $f_0 + f_r$ and the cross-range is perpendicular to the down-range.

The frequency response $A(f_r, \theta, \varphi)$ after discretized to an $M \times N$ matrix is composed of

$$A(f_r, \theta, \varphi) = \begin{pmatrix} A(f_1, \theta, \beta_1) & A(f_2, \theta, \beta_1) & \dots & A(f_N, \theta, \beta_1) \\ A(f_1, \theta, \beta_2) & A(f_2, \theta, \beta_2) & \dots & A(f_N, \theta, \beta_2) \\ \vdots & \vdots & \vdots & \vdots \\ A(f_1, \theta, \beta_M) & A(f_2, \theta, \beta_M) & \dots & A(f_N, \theta, \beta_M), \end{pmatrix} \quad (10)$$

where $\beta_j, j = 1, 2, \dots, M$ is the sampling point of the rotation angle α centered on the squint angle φ, and $f_i, i = 1, 2, \dots, N$ is the sampling point of the frequency series $f_c + f_r$. Electromagnetic scattering calculations can be seen as the projection of the 3D model on the plane formed by the incident angle. And the reflectivity function $\sigma(t_r, \theta, \varphi)$ can be generated with 2D inverse fast Fourier transform (2DIFFT) on $A(f_r, \theta, \varphi)$ when the rotation angle is relatively small. Specifically, to make the RCS features reasonable in SAR imaging, The rotation angle α should be equal to the synthetic angle in SAR imaging, based on which α is represented by

$$\alpha = 0.886 \frac{\lambda}{L_a}, \quad (11)$$

and the interval of the incident angle is

$$\Delta\beta = \frac{\alpha}{N - 1}. \quad (12)$$

Considering the squint angle φ, and the incident angles β_j are denoted by

$$\beta_j = (j - 1) \cdot \Delta\beta + \varphi - \frac{\alpha}{2}. \quad (13)$$

MoM is a suitable approach for addressing electromagnetic scattering problems with arbitrary geometrical shapes in the lower frequency range. However, as the frequency increases, the computational time and memory requirements often surpass the capabilities of available computers. In addition, for a scatterer with large and smooth surfaces, LEPO method can be modelled using large triangular edge elements. This is achieved by incorporating asymptotic solutions to predict the rapid phase dependence of the unknown current distribution. This yields a slowly varying residual function that can be represented by a coarse density of unknowns. This is akin to an arbitrary field incident on a planar surface, where the phase of the induced current on the surface can be approximated from the phase of the incident field [14], [15]. In hybrid scatters, the rough surface and target RCS often exhibit inconsistencies. The surface is typically electrically large and relatively smooth, while the target is usually smaller in size and contains intricate structures. It is challenging to model such disparate electrical structures accurately using one single method. Therefore, by employing MoM-LEPO hybrid method, the regions can be divided and discretized into different scales to reduce the number of elements for the electrically large surfaces. The coupling effects between different regions are then computed, enabling a relatively accurate calculation of the scattering field while significantly reducing the computational time and memory consumption. To demonstrate the above analysis, we have calculated the RCS of a one-meter diameter metal sphere via these methods, respectively. The simulation results are listed as in Table 1.

The RCS of a one-meter diameter metal sphere has a true value of $\pi/4$. Simulation results in Table 1 indicate that the MoM method achieves the highest accuracy but requires significant memory and time resources. On the other hand, the PO method has lower memory requirements and faster computation time but produces considerable errors. By employing a MoM-LEPO hybrid approach, the computational

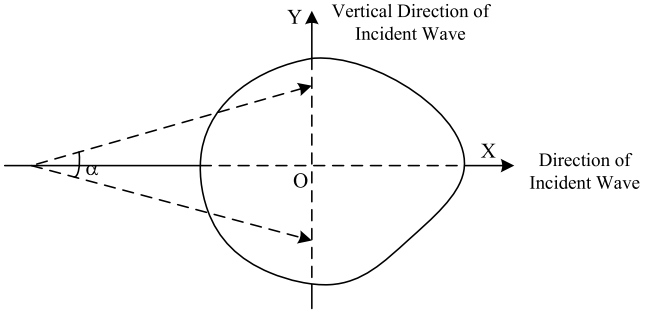


Fig. 2 Sketch map of incident wave direction in electromagnetic calculations.

Table 1 Comparison of different electromagnetic computation methods.

Methods	MoM	PO	MoM-LEPO
RCS /m^2	0.788	0.762	0.769
Error rate	0.3%	3.0%	2.1%
Memory /GB	19.4	0.05	0.06
Time /s	10016.6	3.5	3.8

accuracy can be improved while utilizing lower memory and time resources. This makes it suitable for calculation complex RCS of electrically large objects. Hence, we utilize the hybrid MoM-LEPO method for RCS calculations in this work.

2.3 SAR Image Generation

With the target RCS $A(f_r, \theta, \varphi)$ substituted in (7), raw echo signal in the 2-D frequency domain is simulated. RDA, characteristic of simple and efficient, is adopted to obtain the SAR image. The specific details are referred to [22]. The focused SAR imaging result is given by

$$s_{img}(t_r, t_a, \theta, \varphi) = \sigma(t_r, \theta, \varphi) \cdot sinc\left(\frac{t_r - 2R_0/c}{T}\right) \cdot sinc(B_a t_a), \tag{14}$$

where B_a represents the Doppler bandwidth.

2.4 Overall Flowchart

Based on the above analysis, the procedures are given as in Fig. 3, which illustrates the whole process of the proposed SAR image generation method with fast RCS calculation speed assisted by MoM-LEPO and low computational load via 2DFFT. Basically, there are three steps and the details are as follows.

- Step 1: Raw echo simulation. According to the SAR system parameters, raw echo signal in the 2D frequency domain is generated after performing 2DFFT.
- Step 2: MoM-LEPO-based RCS calculation. To calculate the target RCS, electromagnetic calculation in aid of hybrid MoM-LEPO method is carried out with the frequency points in Step 1 and the derived incident angles.
- Step 3: SAR image acquisition. By substituting the target RCS in Step 2 into the raw echo in the 2D frequency domain in Step 1, the echo reflected from the target in the 2D frequency domain is obtained. With the matched filters derived in RDA, SAR image of the target is acquired.

3. Simulations and Results

3.1 Results and Discussion

Based on the above analysis and fundamentals of SAR imaging [22], experiments are carried out using the system parameters listed in Table 2 to verify the efficiency of the proposed SAR image generation method.

Figure 4 illustrates the 3D generic aeroplane model a380 with a size of 72.44 m × 75.41 m × 23.45 m to perform RCS calculations.

We choose the squint angle $\varphi = 0^{\circ}, 5^{\circ}, 25^{\circ}$ to verify that the proposed method is feasible when SAR works in the side-looking and squint-looking mode. And the relevant parameters are listed as follows. We take $\varphi = 0^{\circ}$ as an example. Under this circumstance, $\alpha = 0.886\lambda/L_a = 3.1728^{\circ}$, $\beta_1 = -1.5864^{\circ}$ and $\beta_N = 1.5864^{\circ}$. With the certain incident angle settings and hybrid MoM-LEPO method, the normalized RCS (NRCS) curve of the aeroplane for $f_c = 4$ GHz

Table 2 System parameters.

Parameters	Values
Depression Angle θ [°]	30
Carrier Frequency f_c [GHz]	4
Chirp Rate μ [Hz/s]	6×10^{13}
Pulse Repetition Time [μs]	2
Chirp Bandwidth [MHz]	120
Initial Range R_0 [km]	20
Velocity v [m/s]	120
Antenna length [m]	1.2
Number of Range Samples	1024
Number of Azimuth Samples	2048
Range Oversampling Rate	1.2
Azimuth Oversampling Rate	1.25
Polarization mode	HH
Normalized linear magnitude	[0,1]

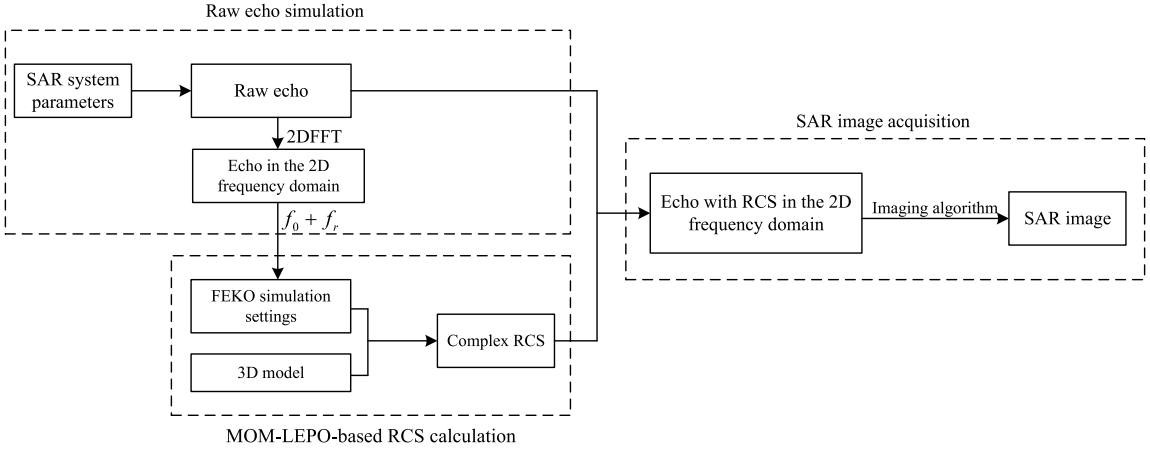


Fig. 3 Flowchart of SAR image generation for a 3D target with consideration of RCS.

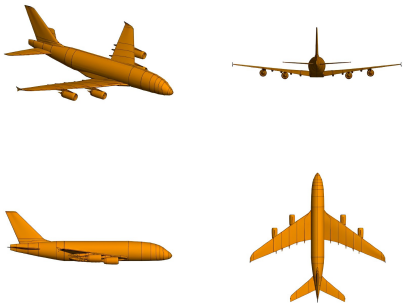


Fig. 4 3D aeroplane model.

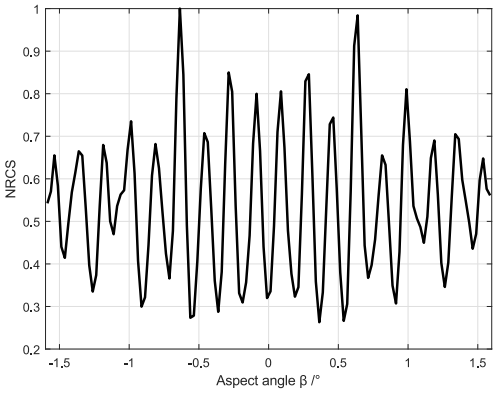


Fig. 5 NRCS vs aspect angle β for $\varphi = 0°$.

is shown in Fig. 5. According to Fig. 5, the RCS curve is symmetric about $\beta = 0°$, which is consistent with the symmetric structure of the aeroplane. Additionally, RCS value over certain incident angle is relatively small. For example, when $\beta = 0°$, the scattering properties are not intense in this direction compared with those in other directions.

As stated in [23], [24], HRRP can be used in approximately estimating the locations and intensities of the scattering points composed by important parts of the target, conducive to target recognition and detection. With the RCS curve in Fig. 5, we perform IFFT on the RCS data when $\beta = 0°$ to obtain high resolution range profile (HRRP) as shown in Fig. 6. Under the incident angle $\theta = 30°, \phi = 0°$, the peaks of HRRP for the aeroplane reveals the existence of crucial parts, such as nose, fuel tanks, landing gear, wings and tail. Furthermore, the distance between peaks and the actual distance between components are proportional. For example, the distance between nose and tail is approximately equal to the product of length of the plane and $\sin\theta$, which means that the HRRP can be viewed as the projection of the 3D target in the incident direction of plane wave.

Considering that HRRP in Fig. 6 displays the key components of the aeroplane along the range axis, the azimuth position of which can be obtained with the inclusion of the rotation angle. Thus, HRRP for the incident angles within the rotation angle α set according to the value of squint angle φ will constitute for RCS of the aeroplane. In other words, by performing 2DIFFT on frequency response $A(f_r, \varphi, \theta)$ denoted as (10), RCS can be obtained, revealing the scat-

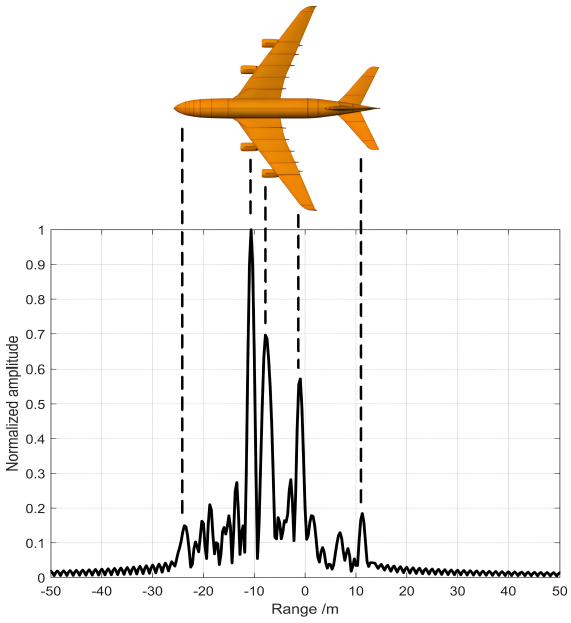
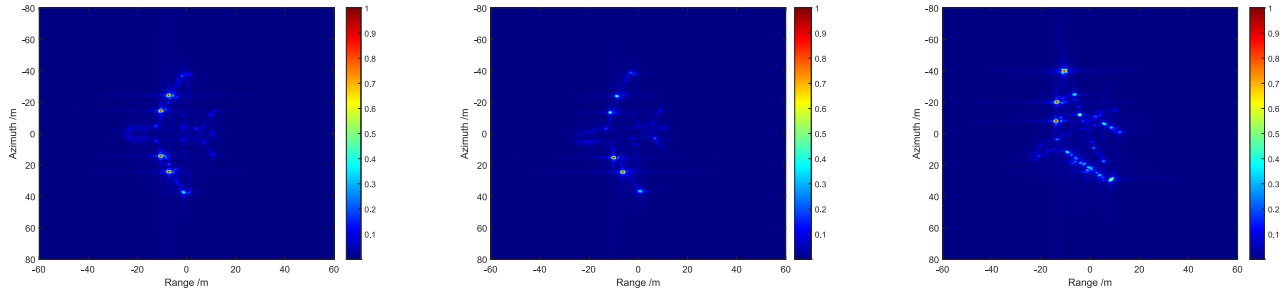


Fig. 6 HRRP of the aeroplane.

tering characteristics of the aeroplane. As shown in Fig. 7, the important components of an aeroplane, such as nose, wings, landing gear and tail, can be vividly seen. Besides, the scattering characteristics on the central axis of the aircraft are not intense in Fig. 7(a), which is consistent with the finding for $\beta = 0°$ in Fig. 5. Furthermore, RCS show that the aeroplane tilts with different incident angles caused by increasing squint angle φ, verifying that RCS calculation similar to turntable imaging can be seen as the projection of the 3D target in the incident direction.

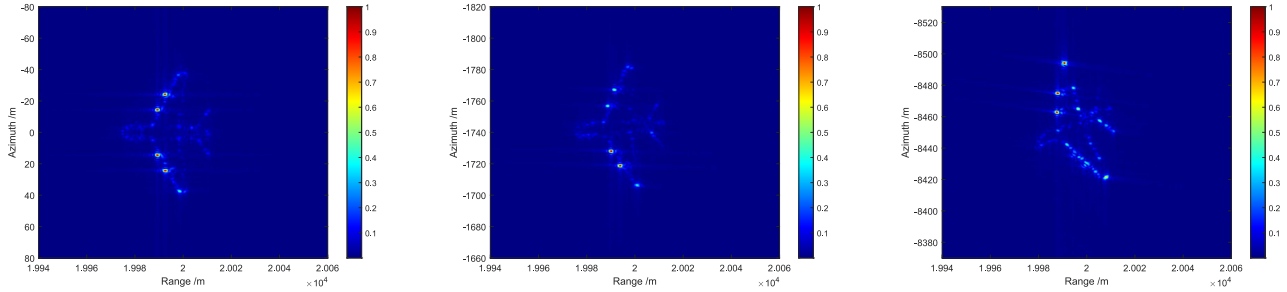
Based on the LFM transmitted signal and target RCS, raw echo signal in the 2-D frequency domain is simulated. By multiplying the matched filters, SAR images for different squint angles are shown in Fig. 8. We observe that the strong scattering points in SAR images are consistent with original RCS, also revealing the crucial parts of the aeroplane model. This means that through the proposed SAR image generation method, SAR image samples for different models and angles can be generated. Furthermore, the aeroplane in the SAR image tilts for a certain squint angle compared with the corresponding RCS.

Considering that the target SAR imaging can be seen as the amplitude and phase modulation of the reference point, we resort to the amplitude spectrum slice by upsampling 32×32 points centered around the reference point by 8 times for different squint angles to explain the tilt of SAR images in Fig. 9. To make a better comparison of RCS and SAR images, the azimuth sidelobes are parallel to the azimuth axis. Since that the Doppler centroid frequency is $2(f_c + f_r)\, v \sin\varphi / c$, the sampling point of azimuth frequency is related to the range frequency f_r. This leads to the tile of the range spectrum, i.e., as the squint angle φ increases, the range sidelobes tilt more away from the horizontal axis.



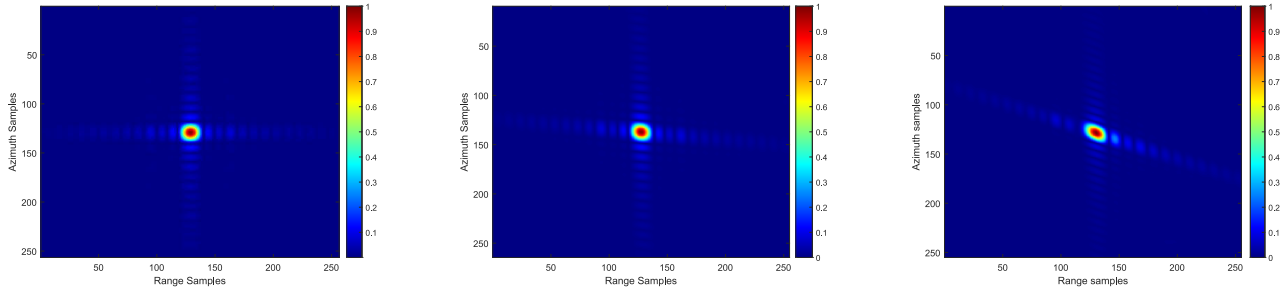
(a) $\varphi = 0°$ (b) $\varphi = 5°$ (c) $\varphi = 25°$

Fig. 7 RCS of the aeroplane for different squint angles.



(a) $\varphi = 0°$ (b) $\varphi = 5°$ (c) $\varphi = 25°$

Fig. 8 SAR image with LFM transmitted signal for different squint angles.



(a) $\varphi = 0°$ (b) $\varphi = 5°$ (c) $\varphi = 25°$

Fig. 9 Amplitude spectrum slice centered around the reference point for different squint angles.

3.2 Computational Efficiency Analysis

According to [7], the computational efficiency of TD method is denoted as

$$C_{TD} = M^2N^2. \tag{15}$$

Comparatively, raw echo simulation based on 1DFFT proposed by [10] and [11] reduces the computational complexity, which can be written as

$$C_{1DFFT} = MNC_R + 1/2MNlog_2(N) + MN \tag{16}$$

where C_R represents the complexity for computing single frequency RCS of the whole target, and $1/2MNlog_2(N)$ denotes IFFT operation. Comparatively, the proposed 2DFFT method omits this step and perform range and azimuth compensation in the frequency domain. Thus, the computational load is reduced to

$$C_{total} = MNC_R + MN. \tag{17}$$

Comparatively, it is evident that the proposed method has a lower computational complexity, which would improve the efficiency of SAR image generation and lays a foundation for waveform design and SAR image interpretation.

4. Conclusion

This work presents a framework of SAR image generation for a 3D target in aid of hybrid MoM-LEPO electromagnetic calculations, which lays the foundation for radar waveform design and SAR image interpretation. Specifically, multiplying raw echo signal after 2DFFT and MoM-LEPO-based RCS yields target echo in the 2D frequency domain, based on which SAR image is generated via RDA. This

method reduces computational load and speed up calculation by performing 2DFFT in raw echo signal and using FEKO solver. Furthermore, simulation results show that the proposed method is applicable in different squint angles.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62371236.

References

[1] H. Huang, F. Gao, J. Wang, A. Hussain, and H. Zhou, "An incremental SAR target recognition framework via memory-augmented weight alignment and enhancement discrimination," IEEE Geosci. Remote Sens. Lett., vol.20, 4005205, 2023.
[2] Y. Yuan, S. Chen, S. Zhang, and H. Zhao, "A chirp scaling algorithm for forward-looking linear-array SAR with constant acceleration," IEEE Geosci. Remote Sens. Lett., vol.15, no.1, pp.88–91, Jan. 2018.
[3] C. Wang, S. Luo, J. Pei, X. Liu, Y. Huang, Y. Zhang, and J. Yang, "An entropy-awareness meta-learning method for SAR open-set ATR," IEEE Geosci. Remote Sens. Lett., vol.20, 4005105, 2023.
[4] R.V. Fonseca, R.G. Negri, A. Pinheiro, and A. Atto, "Wavelet spatio-temporal change detection on multitemporal SAR images," IEEE J. Sel. Topics Appl. Earth Observ., vol.16, pp.4013–4023, 2023.
[5] J. Chen, M. Xing, X.G. Xia, J. Zhang, B. Liang, and D.G. Yang, "SVD-based ambiguity function analysis for nonlinear trajectory SAR," IEEE Trans. Geosci. Remote Sens., vol.59, no.4, pp.3072–3087, April 2021.
[6] R. Hu, B.S.M.R. Rao, A. Murtada, M. Alaee-Kerahroodi, and B. Ottersten, "Automotive squint-forward-looking SAR: High resolution and early warning," IEEE J. Sel. Topics Signal Process., vol.15, no.4, pp.904–912, June 2021.
[7] A. Mori and F. De Vita, "A time-domain raw signal simulator for interferometric SAR," IEEE Trans. Geosci. Remote Sens., vol.42, no.9, pp.1811–1817, Sept. 2004.
[8] Q. Song, F. Xu, and Y.Q. Jin, "SAR image representation learning with adversarial autoencoder networks," IGARSS 2019 - 2019 IEEE Int. Geosci. and Remote Sens. Symp., Yokohama, Japan, pp.9498–9501, IEEE, July 2019.
[9] Q. Song, H. Chen, F. Xu, and T.J. Cui, "EM simulation-aided zero-shot learning for SAR automatic target recognition," IEEE Geosci. Remote Sens. Lett., vol.17, no.6, pp.1092–1096, June 2020.
[10] W. Xia, Y. Qi, L. Huang, and X. Jin, "Missile-borne SAR raw signal simulation for maneuvering target," Int. J. Antennas Propag., vol.2016, pp.1–12, 2016.
[11] C.l. Dong, X. Meng, and L.x. Guo, "Research on SAR imaging simulation based on time-domain shooting and bouncing ray algorithm," IEEE J. Sel. Topics Appl. Earth Observ., vol.16, pp.1519–1530, 2023.
[12] R. Li, K. Ji, H. Zou, and S.L. Zhou, "Simulation of SAR imagery of target based on electromagnetic scattering characteristic computation," Radar Sci. and Technol., vol.22, no.5, pp.395–400, 2010.
[13] T. El-Moselhy and L. Daniel, "Variation-aware stochastic extraction with large parameter dimensionality: Review and comparison of state of the art intrusive and non-intrusive techniques," 2011 12th International Symposium on Quality Electronic Design, pp.1–10, 2011.
[14] N.K. Sahoo, D.C. Panda, and S.K. Dash, "RCS studies of anti-ship missile over sea surface using LE-PO method," 2017 IEEE Appl. Electromagn. Conf. (AEMC), pp.1–2, 2017.
[15] A. Altintas and A. Celik, "Large flat plate models in the physical optics method for RCS calculations," 10th Int. Conf. on Math. Methods in Electromagn Theory, 2004., pp.586–588, 2004.
[16] H. Mohammadzadeh, A.Z. Nezhad, and Z.H. Firouzeh, "Modified physical optics approximation for RCS calculation of dielectric coated PEC with axial symmetry," 2013 21st Iranian Conf. on Elect. Eng. (ICEE), pp.1–5, 2013.
[17] U. Jakobus and F. Landstorfer, "Improved PO-MM hybrid formulation for scattering from three-dimensional perfectly conducting bodies of arbitrary shape," IEEE Trans. Antennas Propag., vol.43, no.2, pp.162–169, 1995.
[18] Q. Chang, C. Sun, and Y. Wang, "RCS measurement method based on high-resolution sparse turntable imaging," 2018 Int. Conf. on Microwave and Millimeter Wave Technology (ICMMT), Chengdu, pp.1–3, IEEE, May 2018.
[19] X. Wang, C. Wang, and Y. Liu, "RCS computation and Analysis of target using FEKO," Proc. of 2014 3rd Asia-Pacific Conf. on Antennas and Propag., Harbin, China, pp.822–825, IEEE, July 2014.
[20] X. Lu, J. Xia, Z. Yin, and W. Chen, "High resolution turntable radar imaging via two dimensional deconvolution with matrix completion," Sensors, vol.17, no.3, p.542, March 2017.
[21] S. Demirci, "Applying polarimetric target decomposition to 2D turntable ISAR imagery of a complex vehicle," Int. J. Antennas Propag., vol.2022, pp.1–15, July 2022.
[22] I.G. Cumming and F.H. Wong, Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation, Artech House, Norwood, 2005.
[23] W. Yue and Y. Xin, "Performance analysis of radar high-resolution range profiling for stationary targets," 2017 IEEE Int. Conf. on Signal Process., Commun. and Comput. (ICSPCC), Xiamen, pp.1–5, IEEE, Oct. 2017.
[24] L. Du, H. Liu, Z. Bao, and M. Xing, "Radar HRRP target recognition based on higher order spectra," IEEE Trans. Signal Process., vol.53, no.7, pp.2359–2368, July 2005.

Xian Yu was born in 1997. She received the B.S. degree in electronic engineering from Nanjing University of Science and Technology, China, in 2019, where she is currently pursuing the Ph.D. degree. Her research interests include SAR imaging, maritime target detection and radar signal processing.

Yubing Han was born in 1971. He received the Ph.D. degree in signal and information processing from Southeast University, China, in 2006. Since 2006, he has been a Faculty Member with the School of Electronic and Optical Engineering in Nanjing University of Science and Technology, where he is currently a Professor. His current research interests include SAR imaging, sensor signal processing, radar signal processing, wireless communications, and digital image processing.