PAPER

NOMA-Based Optimal Multiplexing for Multiple Downlink Service Channels to Maximize Integrated System Throughput*

Teruaki SHIKUMA[†], *Student Member*, Yasuaki YUDA[††], *Member*, *and* Kenichi HIGUCHI[†a)], *Senior Member*

SUMMARY We propose a novel non-orthogonal multiple access (NOMA)-based optimal multiplexing method for multiple downlink service channels to maximize the integrated system throughput. In the fifth generation (5G) mobile communication system, the support of various wireless communication services such as massive machine-type communications (mMTC), ultra-reliable low latency communications (URLLC), and enhanced mobile broadband (eMBB) is expected. These services will serve different numbers of terminals and have different requirements regarding the spectrum efficiency and fairness among terminals. Furthermore, different operators may have different policies regarding the overall spectrum efficiency and fairness among services. Therefore, efficient radio resource allocation is essential during the multiplexing of multiple downlink service channels considering these requirements. The proposed method achieves better system performance than the conventional orthogonal multiple access (OMA)-based multiplexing method thanks to the wider transmission bandwidth per terminal and inter-terminal interference cancellation using a successive interference canceller (SIC). Computer simulation results reveal that the effectiveness of the proposed method is especially significant when the system prioritizes the fairness among terminals (including fairness among services).

key words: *non-orthogonal multiple access, successive interference cancellation, bandwidth allocation, power allocation, system throughput, eMBB, mMTC, URLLC*

1. Introduction

In contrast to the fourth generation systems such as Long Term Evolution (LTE) and LTE-Advanced [1], [2] where the primary service offered is mobile broadband, the fifth generation (5G) system is expected to support a wider range of wireless communication services (use cases) such as massive machine-type communications (mMTC) and ultra-reliable low latency communications (URLLC), in addition to enhanced mobile broadband (eMBB) [3]–[5]. The physical channels dedicated to these wireless communication services will be multiplexed within a shared system frequency band and the respective wireless communication services will have different requirements for spectrum efficiency and fairness among user terminals (machines). For example, mMTC and URLLC are assumed to give priority to the fairness between terminals rather than the spectrum efficiency compared to eMBB. Furthermore, different operators may have different policies regarding the overall spectrum efficiency and fairness among services. Therefore, in order to maximize the performance of the entire system, all the radio resources (bandwidth and power) should be shared appropriately among service channels considering the service requirements, number of terminals, and their channel conditions for all wireless communication services. This paper focuses on the downlink.

Members of our research group recently reported a downlink frequency bandwidth allocation method among service channels [6]–[8]. In these studies, the optimization criterion is to maximize the defined system throughput that integrates the system performance of all services considering the above requirements. The reported method jointly optimizes time/frequency-domain scheduling and frequency bandwidth allocation among service channels in terms of maximizing the integrated system throughput. The reported method can be categorized into an orthogonal multiple access (OMA)-based multiplexing approach.

Non-orthogonal multiple access (NOMA) has gained attention as a candidate multiple access scheme for future radio access [9]–[13]. From an information-theoretic perspective, NOMA with a successive interference canceller (SIC) is an optimal multiple access scheme from the viewpoint of the achievable multiuser capacity region for the downlink, i.e., broadcast channel [14]. The throughput gain when using NOMA compared to that for OMA is especially significant when the system prioritizes the fairness among terminals [9].

We propose a novel NOMA-based optimal multiplexing method for multiple downlink service channels to maximize the integrated system throughput. We derive an optimal scheduling (bandwidth allocation) and power allocation method to the terminals for all services. The proposed method achieves better system performance than the conventional OMA-based multiplexing method [7] thanks to the wider transmission bandwidth per terminal and inter-terminal interference cancellation due to the use of the SIC. Computer simulation results reveal that the effectiveness of the proposed method is especially significant when the system gives priority to the fairness among terminals (including fairness among services). We note that this paper is an extended version of our international conference paper [15] that includes enhanced evaluation and discussions.

Manuscript received November 22, 2019.
Manuscript revised March 16, 2020.
Manuscript publicized May 20, 2020.

[†] The authors are with the Department of Electrical Engineering, Graduate School of Science and Technology, Tokyo University of Science, Noda-shi, 278-8510 Japan.

[††] The author is with Panasonic Corporation, Yokohama-shi, 224-8539 Japan.

[*] The material in this paper was presented in part at the IEEE 90th Vehicular Technology Conference, Honolulu, U.S.A., September 2019.

a) E-mail: higuchik@rs.tus.ac.jp

DOI: 10.1587/transcom.2019EBP3241

The remainder of this paper is organized as follows. First, Sect. 2 presents the system model including the definition of the integrated system throughput to be maximized. Section 3 describes the proposed method. In Sect. 4, the system throughput levels for the proposed method are comparatively evaluated with those of the conventional methods based on computer simulations. Section 5 concludes the paper.

2. System Model

The proposed method is applied at each base station (cell) independently. Hence, we describe the system model at a specific cell of interest in the following. The set of services is denoted as $\mathcal{N}_{\mathrm{S}}$. The set of all terminals that receive service $i \in \mathcal{N}_{\mathrm{S}}$ in the cell is denoted as $\mathcal{K}_i$. The set of all terminals of all services in the cell is denoted as $\mathcal{K} = \cup_{i\in\mathcal{N}_{\mathrm{S}}}\mathcal{K}_i$. The set of frequency blocks in the entire system frequency bandwidth, which should be shared by all the services, is denoted as $\mathcal{F}$.

The performance indicator for each service is represented as the system throughput calculated from the generalized mean [16] of the throughput of all terminals receiving that service. The system throughput of service i at discrete time t, $C_i(t)$, is defined as

$$C_i(t) = \left(\frac{1}{|\mathcal{K}_i|}\sum_{k\in\mathcal{K}_i} R_k(t)^{m_i}\right)^{\frac{1}{m_i}}. \tag{1}$$

Here, $R_k(t)$ is the average throughput of terminal k at time t. Parameter m_i is equal to or less than 1. When m_i is 1, (limit of) 0, -1, and $-\infty$, the system throughput, $C_i\ (t)$, corresponds to the arithmetic mean, geometric mean, harmonic mean, and minimum of the terminal throughput levels within a cell, respectively. The choice of m_i is dependent on the operational policy regarding the tradeoff between the fairness among terminals and the spectrum efficiency of service i. In this paper, we consider the fairness regarding the achievable throughput. For example, when the throughput levels of all terminals are equalized, it is argued that fairness is best achieved. When increasing the fairness among terminals takes priority over the spectrum efficiency, m_i should be set low and vice versa.

We define the integrated system throughput, $C(t)$, by the generalized mean of $\{C_i(t)\}$ as an overall system performance indicator. Term $C(t)$ is represented as

$$C(t) = \left(\frac{1}{|\mathcal{N}_{\mathrm{S}}|}\sum_{i\in\mathcal{N}_{\mathrm{S}}} \{\alpha_i C_i(t)\}^{m_{\mathrm{S}}}\right)^{\frac{1}{m_{\mathrm{S}}}} \tag{2}$$

where α_i ($\alpha_i > 0$) is the priority (weighting) factor for service i and m_{S} is interpreted as a parameter that controls the tradeoff between the overall spectrum efficiency of all services and the fairness among services. As m_{S} is set lower, $C(t)$ takes higher priority in terms of the fairness among services.

3. Proposed Method

The channel gain normalized by the noise power including inter-cell interference of terminal k at frequency block f at time t is denoted as $g_{k,f}(t)$. We assume that the scheduler allocates frequency block f to a set of terminals, $\mathcal{U}_f(t) \subseteq \mathcal{K}$, at time t. We note that $|\mathcal{U}_f(t)|$ is one in the OMA case, while $|\mathcal{U}_f(t)|$ can be greater than one in the NOMA case. The allocated transmission power to terminal $k \in \mathcal{U}_f(t)$ at t is $p_{k,f}(t)$. We assume that the transmission power of each frequency block is limited to p_{total}. Therefore, the transmission power allocation constraint is represented as

$$p_{k,f}(t) \geq 0, \qquad \sum_{k\in\mathcal{U}_f(t)} p_{k,f}(t) = p_{\mathrm{total}}. \tag{3}$$

The set of $p_{k,f}(t)$ of all scheduled terminals at frequency block f at time t is denoted as $\mathcal{P}_f(t)$. The purpose of the proposed method is to optimize $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$ jointly to maximize integrated system throughput $C(t)$.

We assume that the SIC is applied at the terminal receiver to remove the inter-terminal interference. With the SIC, the order of decoding should be in the order of the increasing channel gain normalized by the noise and inter-cell interference power, $g_{k,f}(t)$ [9], [14]. Based on this order, any terminal can correctly decode and cancel the signals of other terminals whose decoding order comes before that terminal. Thus, terminal k can remove the inter-terminal interference from terminal j whose $g_{j,f}(t)$ is lower than $g_{k,f}(t)$. Therefore, the instantaneous throughput of terminal k at frequency block f assuming that the scheduler schedules terminal set $\mathcal{U}_f(t)$ and allocated transmission power set $\mathcal{P}_f(t)$, $r_{k,f}(\mathcal{U}_f(t), \mathcal{P}_f(t); t)$, is represented as

$$r_{k,f}\left(\mathcal{U}_f(t), \mathcal{P}_f(t); t\right) = \begin{cases} W\log_2\left(1+\dfrac{g_{k,f}(t)p_{k,f}(t)}{\displaystyle\sum_{j\in\mathcal{U}_f(t), g_{k,f}(t)<g_{j,f}(t)} g_{k,f}(t)p_{j,f}(t)+1}\right), & k\in\mathcal{U}_f(t) \\ 0, & k\notin\mathcal{U}_f(t) \end{cases}. \tag{4}$$

Here, W is the transmission bandwidth of one frequency block.

The average throughput of terminal k at time t, $R_k(t)$, is defined based on the exponential moving average with the averaging window size of T_{avg} as

$$R_k(t) = R_k(t-1) + \frac{1}{T_{\mathrm{avg}}}\left[\sum_{f\in\mathcal{F}} r_{k,f}(\mathcal{U}_f(t), \mathcal{P}_f(t); t) - R_k(t-1)\right]. \tag{5}$$

From (1), (2), and (5), based on the Taylor expansion, the increase in $C(t)$, $\Delta C(t) = C(t) - C(t-1)$, is represented as

$$\begin{aligned}\Delta C(t) &= \sum_{k\in\mathcal{K}} \frac{\partial C(t-1)}{\partial R_k}[R_k(t)-R_k(t-1)] + O\left(\frac{1}{T_{\text{avg}}^2}\right)\\ &= \frac{1}{T_{\text{avg}}}\sum_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}} \frac{\partial C(t-1)}{\partial R_k} r_{k,f}(\mathcal{U}_f(t),\mathcal{P}_f(t);t)\\ &\quad - \frac{1}{T_{\text{avg}}}\sum_{k\in\mathcal{K}} \frac{\partial C(t-1)}{\partial R_k} R_k(t-1) + O\left(\frac{1}{T_{\text{avg}}^2}\right).\end{aligned} \tag{6}$$

In (6), the second term on the right hand side is a constant that is not a function of $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$ to be determined, and the third term can be ignored when T_{avg} is sufficiently large.

In order to maximize $C(t)$ by maximizing $\Delta C(t)$, $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$ are determined so that the following metric, $\rho_f(\mathcal{U},\mathcal{P};t)$, which corresponds to the first term in (6) at each f after removing constant values, is maximized.

$$\rho_f(\mathcal{U},\mathcal{P};t) = \sum_{k\in\mathcal{K}} \frac{C_{i_k}(t-1)^{m_{\text{S}}} R_k(t-1)^{m_{i_k}-1}}{\sum_{j\in\mathcal{K}_{i_k}} R_j(t-1)^{m_{i_k}}} r_{k,f}(\mathcal{U},\mathcal{P};t). \tag{7}$$

$$(\mathcal{U}_f(t),\mathcal{P}_f(t)) = \arg\max_{(\mathcal{U},\mathcal{P})} \rho_f(\mathcal{U},\mathcal{P};t). \tag{8}$$

In the following, we describe the method to obtain the optimal set of $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$ for each frequency block f at time t. Metric $\rho_f(\mathcal{U},\mathcal{P};t)$ in (7) can be seen as a weighted sum of instantaneous terminal throughput $r_{k,f}(\mathcal{U},\mathcal{P};t)$ where the weighting factor for terminal k, $w_{k,f}(t)$, is

$$w_{k,f}(t) = \frac{C_{i_k}(t-1)^{m_{\text{S}}} R_k(t-1)^{m_{i_k}-1}}{\sum_{j\in\mathcal{K}_{i_k}} R_j(t-1)^{m_{i_k}}}. \tag{9}$$

Therefore, for given candidate scheduling policy $\mathcal{U}$, the metric can be maximized using the power allocation that maximizes the weighted sum of the instantaneous terminal throughput. In [11] and [17], the iterative water-filling power allocation algorithm that achieves this maximization is presented. For the problem at hand, the optimal power allocation algorithm for each frequency block f at time t can be described as follows.

Step 1) Initial setting

- Terminals in given candidate set $\mathcal{U}$ are sorted in decreasing order of weighting factor $w_{k,f}(t)$. The k-th sorted terminal index is denoted as $\pi(k)$.
- $\Delta_f(k) := w_{\pi(k),f}(t) - w_{\pi(k+1),f}(t)$, where $w_{\pi(|\mathcal{U}|+1),f}(t)$ is assumed to be zero.
- $q_f^{(0)}(\pi(k)) := 0$ for $k = 1,\ldots,|\mathcal{U}|$. Term $q_f(\pi(k))$ represents the transmission power for terminal $\pi(k)$ at frequency block f in the dual uplink multiple access channel (MAC).
- Iteration index n is set to 1.

Step 2) Water-filling step for updating power calculation

- $\beta_f^{(n)}(\pi(k),l)$ for all k and $l = 1,\cdots,|\mathcal{U}|$ is calculated as

$$\beta_f^{(n)}(\pi(k),l) := \frac{g_{\pi(k),f}(t)}{1+\sum_{i=1,i\neq k}^{l} g_{\pi(i),f}(t) q_f^{(n-1)}(\pi(i))}. \tag{10}$$

- Updating power $\gamma_f^{(n)}(\pi(k))$ for all k is determined so that

$$\begin{aligned}\nu_k &= \sum_{l=k}^{|\mathcal{U}|} \frac{\Delta_f(l)}{\gamma_f^{(n)}(\pi(k)) + 1/\beta_f^{(n)}(\pi(k),l)} = \mu \quad \forall k\in\mathcal{U},\\ &\text{subject to} \sum_{\pi(k)\in\mathcal{U}} \gamma_f^{(n)}(\pi(k)) = p_{\text{total}}.\end{aligned} \tag{11}$$

Step 3) Update of the transmission power in the dual MAC

- $q_f^{(n)}(\pi(k))$ for all k is updated as

$$q_f^{(n)}(\pi(k)) := \frac{1}{|\mathcal{U}|}\gamma_f^{(n)}(\pi(k)) + \left(1-\frac{1}{|\mathcal{U}|}\right) q_f^{(n-1)}(\pi(k)). \tag{12}$$

Step 4) $n := n+1$. Return to Step 2 for sufficient convergence.

After convergence, the set of $\{q_f^{(n)}(k)\}$ for the dual MAC is converted to power allocation $\mathcal{P} = \{p_{k,f}\}$ in the downlink based on the uplink-downlink duality presented, e.g., in [18].

In the proposed method, at each frequency block, $\mathcal{P}$ is optimized using the iterative water-filling power allocation algorithm above for each candidate $\mathcal{U}$, and the corresponding metric in (7) is evaluated. After evaluating all candidates of $\mathcal{U}$, the set of $\mathcal{U}$ and $\mathcal{P}$ that has the highest metric value is selected as $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$.

The proposed power allocation based on the maximization of the weighted sum of the instantaneous terminal throughput yields an additional merit when the channel coding block is distributed over multiple frequency blocks as in LTE [1] and 5G NR [5]. In the downlink, the decoding order of terminals in the SIC should be in the order of the increasing normalized channel gain. This order may change among frequency blocks due to frequency-selective fading. In such a case, when the channel coding block is distributed over multiple frequency blocks, SIC decoding cannot be performed. This problem is addressed by the fact that a terminal is allocated power only if all terminals with larger weighting factors have lower channel gains in the proposed optimal power allocation. Since weighting factor $w_{k,f}(t)$ in (9) for every terminal k is common to all frequency blocks (thus, not a function of f), the decoding order of any pair of terminals, to which the power is allocated, becomes the same for all frequency blocks. Actually, the decoding order of terminals is equivalent to the order of the decreasing weighting factor, $w_{k,f}(t)$ [19]. We also note that when we use a sub-optimal but less-complex power allocation algorithm than the proposed optimal method, the mismatch in decoding order of terminals in the SIC among frequency blocks

Table 1 Simulation parameters.

System bandwidth		9 MHz
Number of frequency blocks		50
Number of service channels		2
Node density	Base station	1 / km^2
	Terminal	$\{\lambda_1, \lambda_2\}$ = {25, 25}, {40, 10} / km^2
Transmission power of base station		46 dBm
Distance-dependent path loss (including antenna gain)		114.1+37.6log$_{10}$(r), r: kilometers
Shadowing		Lognormal shadowing with standard deviation of 8 dB and inter-site correlation of 0.5
Instantaneous fading		Six-path Rayleigh, rms delay spread = 1 µs and f_D = 5.55 Hz
Receiver noise power density (including noise figure)		−165 dBm/Hz
Scheduling interval		1 ms
Maximum number of non-orthogonally multiplexed terminals		2 per frequency block

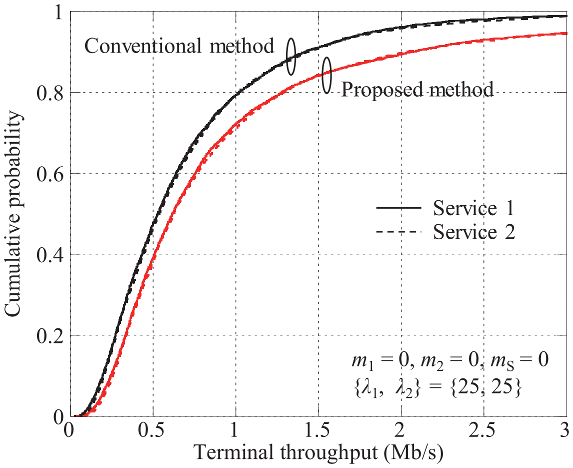

Fig. 1 Distribution of terminal throughput (when m_1 = 0).

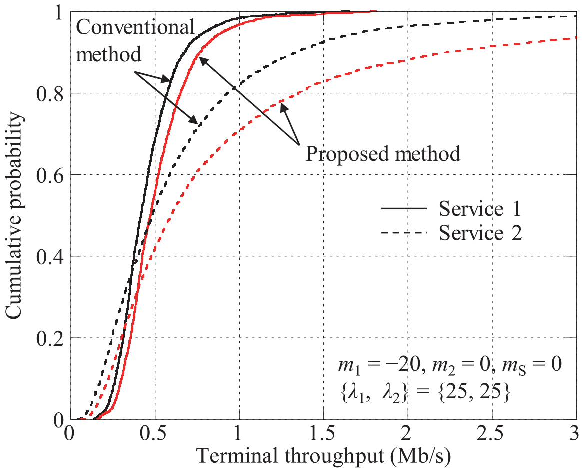

Fig. 2 Distribution of terminal throughput (when m_1 = −20).

must be solved through unified terminal grouping, e.g., in [20].

4. Numerical Results

The system throughput performance of the proposed method is evaluated based on computer simulations. Table 1 gives the simulation parameters. We assume a 9-MHz system bandwidth with 50 frequency blocks ($|\mathcal{F}| = 50$) with a bandwidth of 180 kHz. Universal frequency reuse is assumed among cells. We assume that there are two service channels that are multiplexed within a system bandwidth, and the priority factors, $\{\alpha_i\}$, for all services that are used during the calculation of integrated system throughput $C(t)$ in (2) are set to one for simplicity. The base stations and terminals of each service are placed in random locations within a wrap-around 5 × 5-square kilometer system coverage area based on the Poisson point process (PPP). The node density of base stations is set to 1 per square kilometer. The node densities of the terminals per square kilometer in services 1 and 2 are denoted as λ_1 and λ_2, respectively, which are parameterized in the evaluation. Each terminal is assumed to be associated with the base station whose received power is the maximum.

The transmission power level of the base station is 46 dBm. As the propagation model, distance dependent path loss; lognormally distributed random shadowing with the standard deviation of 8 dB and inter-site correlation of 0.5; and 6-path Rayleigh fading based on a 6-tapped delay line model with the rms delay spread of 1 μs are simulated. We assume a pedestrian environment and the maximum Doppler frequency, f_D, is set to 5.55 Hz, which corresponds to the terminal mobility of 3 km/h at the carrier frequency of 2 GHz. The receiver noise power density of the terminal including the noise figure at the terminal receiver is set to −165 dBm/Hz.

The time-frequency-domain scheduling and power allocation interval is 1 ms. The link-level throughput is calculated based on the Shannon formula and the terminal throughput is defined as the 100-ms averaged throughput (thus, $T_{\mathrm{avg}} = 100$). The performance of the proposed method is verified using parameters m_i and m_S in the definition of the system throughput. The maximum number of non-orthogonally multiplexed terminals per frequency block is set to 2, which should be sufficient to obtain the most from the potential gain of NOMA based on [9] and [11]. In addition to the proposed NOMA-based multiplexing method, the conventional OMA-based multiplexing method in [7] is evaluated for comparison.

Figures 1 and 2 show the cumulative probability of the terminal throughput when $\{\lambda_1, \lambda_2\} = \{25, 25\}$ and $m_2 = m_S = 0$. Figure 1 assumes $m_1 = 0$ while Fig. 2 assumes $m_1 = -20$. Thus, we can say that Fig. 2 shows the performance when the system gives higher priority to the fairness among terminals in service 1 than that in Fig. 1. From the figures, the proposed NOMA-based multiplexing method achieves better throughput than the conventional OMA-based method for the entire region of the cumulative distribution. This is because the terminal throughput assuming OMA-based multiplexing is severely limited by the orthogonal bandwidth allocation, which reduces the bandwidth for the respective terminals. NOMA-based multiplex-

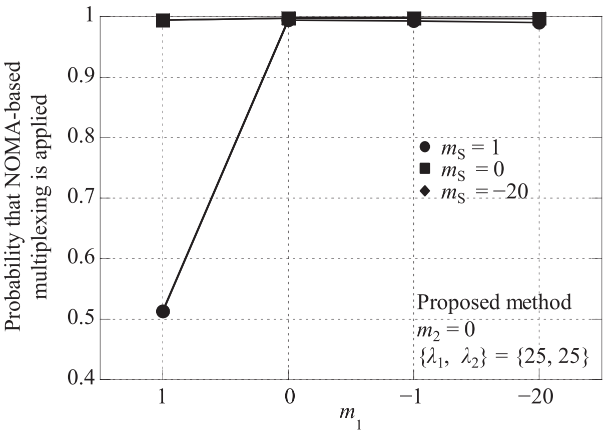

Fig. 3 Probability that NOMA-based multiplexing is applied.

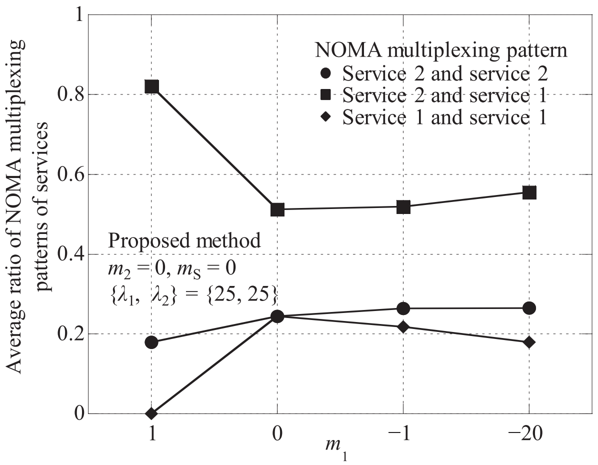

Fig. 4 Average ratio of NOMA multiplexing patterns of services as a function of m_1 (when $m_2 = 0$, $m_S = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

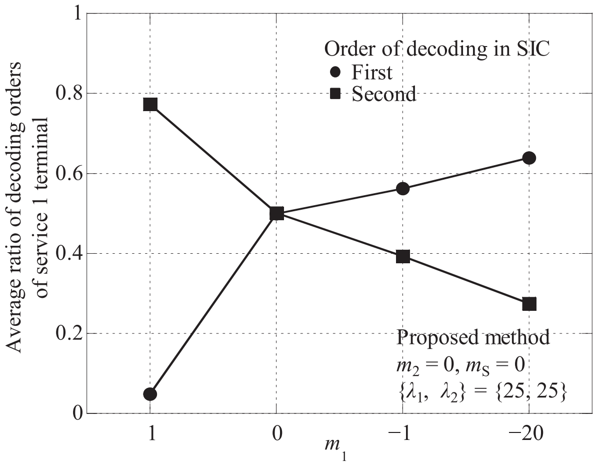

Fig. 5 Average ratio of decoding orders of service 1 terminal as a function of m_1 (when $m_2 = 0$, $m_S = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

ing allows for wider bandwidth usage of all terminals irrespective of their channel conditions. Allocating high power to the power-limited cell-edge terminals associated with the SIC process, which is applied to the bandwidth-limited cell-interior terminals, enhances the throughput of the terminals experiencing a wide range of channel conditions.

From Fig. 2, when the fairness among terminals in service 1 is given higher priority than in Fig. 1, the performance gain when using the proposed method becomes more significant. More specifically, the proposed method avoids significant throughput reduction for the terminals in service 2 compared to the conventional method, in order to achieve the fairness among terminals in service 1. When the fairness among terminals is given higher priority in service 1, the scheduler needs to allocate a significant fraction of bandwidth to the service 1 terminals experiencing poor channel conditions to improve their throughput levels. In conventional OMA-based multiplexing, this results in significant performance degradation for the terminals experiencing good channel conditions or the service giving higher priority to the spectrum efficiency, since the bandwidth allocation to the terminals experiencing good channel conditions is severely limited. When using the proposed NOMA-based multiplexing, all terminals can enjoy a wide transmission bandwidth and inter-terminal interference is effectively mitigated by the appropriate power allocation and SIC.

Figure 3 shows the probability that NOMA-based multiplexing is applied as a function of m_1. Node densities $\{\lambda_1, \lambda_2\}$ are set to $\{25, 25\}$ and m_2 is 0 in the figure. Term m_S is parameterized. When m_1 and m_S are 1, the probability of NOMA-based multiplexing is relatively low. When $m_1 = m_S = 1$, the system heavily favors giving priority to the spectrum efficiency over the fairness among terminals/services. Since the sum capacity (in other words, the arithmetic mean of the terminal throughput) is achieved not only by NOMA with the SIC but also by OMA, the probability that NOMA-based multiplexing is applied seems to be low when $m_1 = m_S = 1$. However, when the other scenario evaluated in Fig. 3 in which the system gives priority to the fairness among terminals/services, the probability that NOMA-based multiplexing is applied is near 100% to achieve better integrated system throughput.

Figure 4 shows the average ratio of the NOMA multiplexing patterns of services as a function of m_1 when NOMA-based multiplexing is applied. Figure 5 shows the average ratio of the decoding order in the SIC of the service 1 terminal as a function of m_1. Node densities $\{\lambda_1, \lambda_2\}$ are set to $\{25, 25\}$ and both of m_2 and m_S are set to 0. From Fig. 4, when m_1 is 1, the NOMA combination of two service 1 terminals is not observed. This is because when $m_1 = 1$, the system gives higher priority to the spectrum efficiency in service 1 and the sum capacity is achieved even with OMA. However, the overall system performance guarantees the fairness among services. Therefore, the NOMA combination of service 1 and service 2 is observed even when m_1 is 1. Figure 4 reveals that as m_1 is decreased from 1, more frequently the NOMA multiplexing of two service 1 terminals is performed.

From Fig. 5, as the system gives higher priority to the fairness among terminals in service 1 than that in service 2

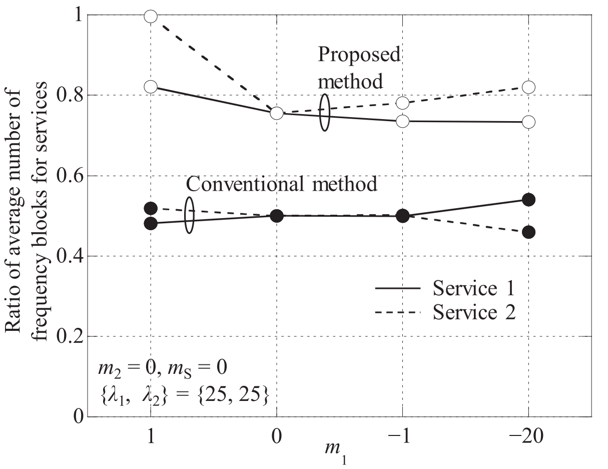

Fig. 6 Ratio of average number of frequency blocks for services as a function of m_1 (when $m_2 = 0$, $m_S = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

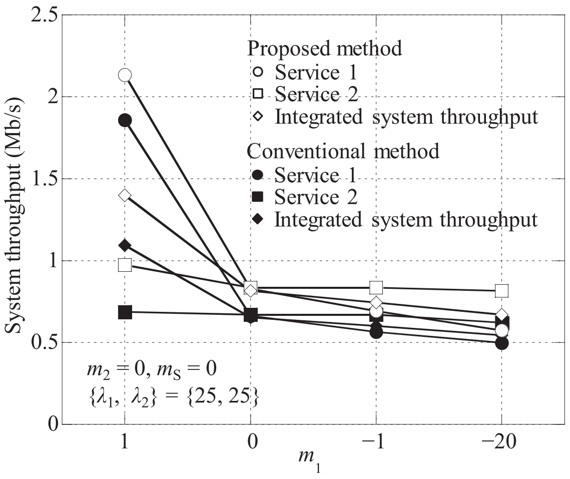

Fig. 8 System throughput as a function of m_1 (when $m_2 = 0$, $m_S = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

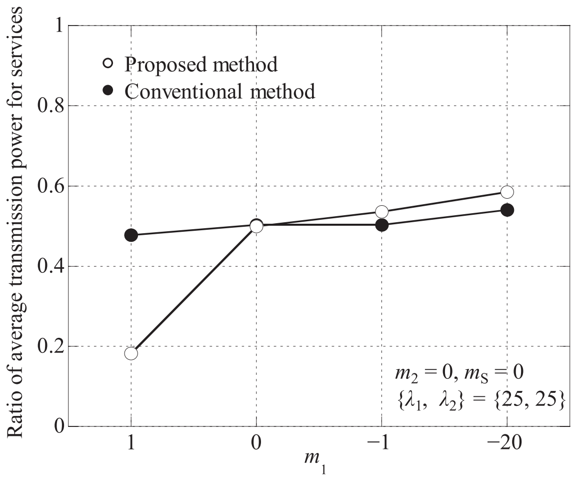

Fig. 7 Ratio of average transmission power for service 1 as a function of m_1 (when $m_2 = 0$, $m_S = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

(in other words, $m_1 < 0$), the terminal in service 1 tends to be decoded first in the SIC process. This is because the terminals in service 1 experiencing relatively poor channel conditions are allocated more radio resources to maintain the fairness among terminals in service 1 and the decoding order of the SIC is in the order of the increasing normalized channel gain.

Figures 6 and 7 show the ratio of the average number of frequency blocks allocated to the respective services and the ratio of the average transmission power allocated to service 1, respectively, as a function of m_1. The ratio of the average number of frequency blocks for each service i is defined as the average of the number of frequency blocks allocated to the terminals receiving service i divided by the total number of frequency blocks (= 50). When using the OMA-based conventional method, the sum of the ratios of the average number of frequency blocks for services 1 and 2 is always 1. When using the NOMA-based proposed method, one frequency block can be used simultaneously by 2 terminals receiving different services and the sum of the ratios of the average number of frequency blocks for services 1 and 2 exceeds 1 in this case. Node densities $\{\lambda_1, \lambda_2\}$ are set to $\{25, 25\}$ and m_2 and m_S are set to 0.

From Fig. 6, the proposed method allocates more frequency blocks to both services compared to the conventional OMA-based method thanks to the NOMA-based multiplexing. When m_1 is 1, the proposed method allocates more radio resources to service 2 than service 1. This is because in this case, only terminals experiencing good channel conditions (in other words, bandwidth-limited cell-interior terminals) in service 1 are allowed to be scheduled. Therefore, the ratio of the average transmission power for service 1 is controlled to be low for better overall power utilization. As m_1 is decreased, more transmission power is allocated to the terminals in service 1 so that the fairness among terminals in service 1 is improved by allocating high power to the terminals experiencing poor channel conditions (thus, power-limited terminals). At the same time, as m_1 is decreased from $m_2 = 0$, more frequency blocks are allocated to the terminals in service 2 to improve the throughput of bandwidth-limited terminals in service 2. In summary, the proposed NOMA-based multiplexing method adaptively and appropriately allocates bandwidth and power resources to the respective service channels depending on the system policy regarding the spectrum efficiency and fairness, which is represented in the definition of the integrated system throughput.

Figure 8 shows the system throughput as a function of m_1 when $m_2 = 0$, $m_S = 0$, and $\{\lambda_1, \lambda_2\} = \{25, 25\}$. Figure 8 shows that the proposed method increases the integrated system throughput for any m_1 compared to that for the conventional method thanks to the wider transmission bandwidth per terminal and inter-terminal interference cancellation using the SIC.

Figures 9 and 10 show the system throughput as a function of m_S when $m_1 = -1$ and $m_2 = 0$, for $\{\lambda_1, \lambda_2\}$ of $\{25, 25\}$ and $\{40, 10\}$, respectively. Both figures show that the pro-

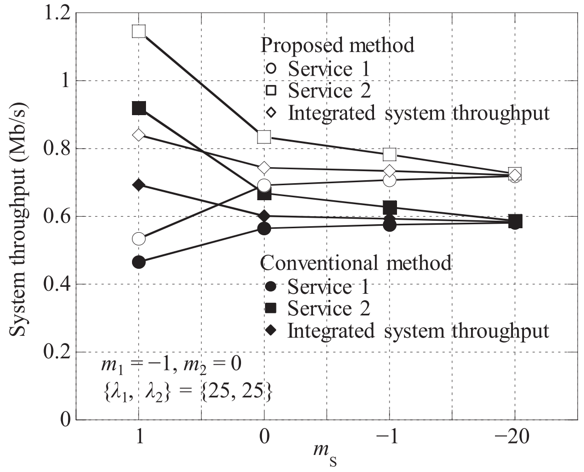

Fig. 9 System throughput as a function of m_S (when $m_1 = -1$, $m_2 = 0$, $\{\lambda_1, \lambda_2\} = \{25, 25\}$).

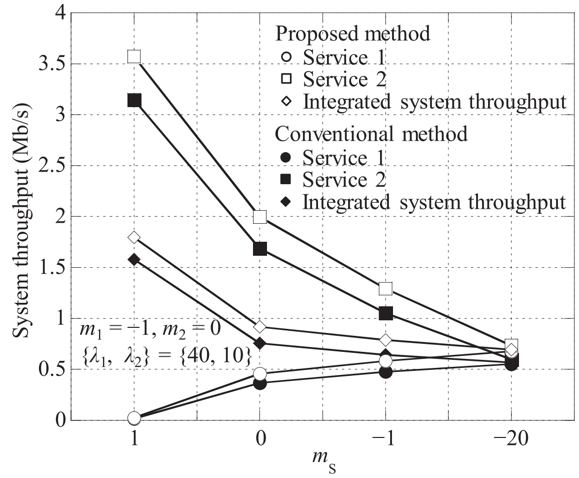

Fig. 10 System throughput as a function of m_S (when $m_1 = -1$, $m_2 = 0$, $\{\lambda_1, \lambda_2\} = \{40, 10\}$).

posed method increases the integrated system throughput for any m_S compared to that for the conventional method thanks to the wider transmission bandwidth per terminal and inter-terminal interference cancellation using the SIC. Figure 9 reveals that the effectiveness of the proposed method is especially significant when m_S is set to a lower value, i.e., when the system gives priority to the fairness among services. This is due to the previously-mentioned NOMA effect. When comparing Figs. 9 and 10, it seems that the performance gain when using the proposed NOMA-based multiplexing compared to that for conventional OMA-based multiplexing is more significant when the numbers of terminals for all services are balanced. However, even when $\{\lambda_1, \lambda_2\}$ is $\{40, 10\}$, a clear increase in the integrated system throughput when using the proposed method for any m_S is confirmed.

5. Conclusion

In this paper, we proposed a novel NOMA-based optimal multiplexing method for multiple downlink service channels to maximize the integrated system throughput. We derived an optimal scheduling (bandwidth allocation) and power allocation method for the terminals of all the considered services. Computer simulation results showed that the proposed method achieves better system performance than that for the conventional OMA-based multiplexing method thanks to the wider transmission bandwidth per terminal and inter-terminal interference cancellation afforded by the SIC. The effectiveness of the proposed method is particularly noteworthy when the system gives priority to the fairness among terminals including fairness among services. It is clear that the wide range of wireless communication services assumed in the 5G mobile communication system such as mMTC, URLLC, and eMBB requires very different requirements in terms of spectrum efficiency and fairness (or per terminal QoS/QoE). The proposed method accommodates these diverse requirements among services and fully utilizes the radio resources (time/frequency bandwidth and transmission power) to achieve optimal system performance.

The system performance gain using the proposed NOMA-based method compared to the OMA-base method in a real system will be affected by the impact of the channel estimation error and the required control signaling overhead. Reference [21] reveals that the performance gain of NOMA with the SIC over OMA in the downlink is not sensitive to the channel estimation error. This is mainly due to the fact that the SIC process, in which channel estimation error results in residual interference after interference cancellation, is applied to a terminal experiencing relatively good channel conditions whose channel estimation accuracy is better than that at the destination terminal of an interfering signal. The control signaling overhead is in general proportional to the number of non-orthogonally multiplexed terminals per channel. The loss in the achievable gain using NOMA due to increased control signaling overhead can be minimized by limiting the number of non-orthogonally multiplexed terminals per channel to two, which is assumed in this paper. We note that in the 3GPP, the technology component option of NOMA (referred to as downlink Multi-User Superposition Transmission (MUST)) was standardized for further enhancement of LTE-Advanced [22], where the number of non-orthogonally multiplexed terminals per channel (resource block) is limited to two. The quantitative performance assessment of the proposed method with more realistic system assumptions is left for future study.

References

[1] 3GPP TR36.913 (V8.0.0), “Requirements for further advancements for E-UTRA (LTE-Advanced),” June 2008.

[2] 3GPP TR36.814 (V9.0.0), “Further advancements for E-UTRA physical layer aspects,” March 2010.

[3] ITU-R M.2083-0, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Sept. 2015.

[4] 3GPP TR38.913 (V0.4.0), "Study on scenarios and requirements for next generation access technologies (Release 14)," June 2016.

[5] 3GPP TR38.912 (V0.0.1), "Study on new radio access technology (Release 14)," June 2016.

[6] S. Mizuno, D. Muramatsu, Y. Yuda, and K. Higuchi, "Investigations on optimum frequency bandwidth allocation method among service channels for system throughput maximization," Proc. APCC2017, Perth, Australia, Dec. 2017.

[7] T. Sakai, Y. Yuda, and K. Higuchi, "Channel-dependent dynamic frequency bandwidth allocation method among service channels to maximize integrated system throughput," Proc. VTC2018-Fall, Chicago, U.S.A., Aug. 2018.

[8] T. Sakai, Y. Yuda, and K. Higuchi, "Inter-base station cooperative scheduling method among multiple service channels to maximize integrated system throughput," Proc. WPMC2018, Chiang Rai, Thailand, Nov. 2018.

[9] 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.

[10] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," IEEE Commun. Mag., vol.53, no.9, pp.74–81, Sept. 2015.

[11] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal multiple access with SIC in cellular downlink using proportional fair-based resource allocation," IEICE Trans. Commun., vol.E98-B, no.2, pp.344–351, Feb. 2015.

[12] N. Nonaka, Y. Kishiyama, and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," IEICE Trans. Commun., vol.E98-B, no.8, pp.1651–1659, Aug. 2015.

[13] K. Higuchi and Y. Kishiyama, "Non-orthogonal multiple access using intra-beam superposition coding and successive interference cancellation for cellular MIMO downlink," IEICE Trans. Commun., vol.E98-B, no.9, pp.1888–1895, Sept. 2015.

[14] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[15] T. Shikuma, Y. Yuda, and K. Higuchi, "NOMA-based optimal multiplexing method for downlink service channels to maximize integrated system throughput," Proc. IEEE VTC2019-Fall, Honolulu, U.S.A., Sept. 2019.

[16] P.S. Bullen, Handbook of Means and Their Inequalities, Kluwer, 2003.

[17] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," IEEE J. Sel. Areas. Commun., vol.24, no.8, pp.1640–1646, Aug. 2006.

[18] P. Viswanath and D.N.C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," IEEE Trans. Inf. Theory, vol.49, no.8, pp.1912–1921, Aug. 2003.

[19] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates and sum rate capacity of Gaussian MIMO broadcast channel," IEEE Trans. Inf. Theory, vol.49, no.10, pp.2658–2668, Oct. 2003.

[20] K. Yakou and K. Higuchi, "Downlink NOMA with SIC using unified user grouping for non-orthogonal user multiplexing and decoding order," Proc. ISPACS 2015, Bali, Indonesia, Nov. 2015.

[21] N. Nonaka, A. Benjebbour, and K. Higuchi, "System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists," Proc. IEEE ICCS2014, Macau, Nov. 2014.

[22] 3GPP TR36.859 (V13.0.0), "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," Dec. 2015.

Teruaki Shikuma received the B.E. degree from Tokyo University of Science, Noda, Japan in 2019. He is currently working towards his M.E. degree in the Department of Electrical Engineering, Tokyo University of Science, Noda, Japan. His research interest includes wireless communications. He is a student member of the IEICE and IEEE.

Yasuaki Yuda received the B.E. and M.E. degrees from Tokyo University of Science, Japan in 1997 and 1999, respectively. He received the Ph.D. degree from Tokai University, Japan, in 2014. Since 1999, he has been with the Matsushita Electric Industrial Co., Ltd. and Panasonic Corporation, Japan. His interests are research and development of wireless communication systems.

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, INC., 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, 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, 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 member of the IEEE.