PAPER

NOMA-Based Highly-Efficient Low-Latency HARQ with Inter-Base Station Cooperation for URLLC

Ryota KOBAYASHI[†], *Nonmember*, **Takanori HARA**[†], **Yasuaki YUDA**[††], *Members*, *and* **Kenichi HIGUCHI**[†a)], *Senior Member*

SUMMARY This paper extends our previously reported non-orthogonal multiple access (NOMA)-based highly-efficient and low-latency hybrid automatic repeat request (HARQ) method for ultra-reliable low latency communications (URLLC) to the case with inter-base station cooperation. In the proposed method, delay-sensitive URLLC packets are preferentially multiplexed with best-effort enhanced mobile broadband (eMBB) packets in the same channel using superposition coding to reduce the transmission latency of the URLLC packet while alleviating the throughput loss in eMBB. Although data transmission to the URLLC terminal is conducted by multiple base stations based on inter-base station cooperation, the proposed method allocates radio resources to URLLC terminals which include scheduling (bandwidth allocation) and power allocation at each base station independently to achieve the short transmission latency required for URLLC. To avoid excessive radio resource assignment to URLLC terminals due to independent resource assignment at each base station, which may result in throughput degradation in eMBB terminals, we employ an adaptive path-loss-dependent weighting approach in the scheduling-metric calculation. This achieves appropriate radio resource assignment to URLLC terminals while reducing the packet error rate (PER) and transmission delay time thanks to the inter-base station cooperation. We show that the proposed method significantly improves the overall performance of the system that provides simultaneous eMBB and URLLC services.

key words: *Hybrid ARQ, non-orthogonal multiple access, NOMA, Inter-base station cooperation, scheduling, transmission rate control, URLLC, eMBB*

1. Introduction

In contrast to 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 New Radio (5G NR) system [3] and beyond [4] are expected to support a wider range of wireless communication services (use cases) such as massive machine-type communications (mMTC), ultra-reliable low latency communications (URLLC), and enhanced mobile broadband (eMBB). This paper focuses on the downlink cellular system in which eMBB and URLLC coexist and investigates a method for improving the low packet error rate (PER) and low transmission latency characteristics of URLLC while suppressing the deterioration in the system throughput for eMBB.

The hybrid automatic repeat request (HARQ) protocol [5] using error detection coding and powerful error correction coding such as the turbo code and low density parity check (LDPC) code efficiently achieves low PER transmission. Therefore, the effective use of HARQ is promising in efficiently achieving a low PER for URLLC. However, increased transmission latency when packet retransmission is conducted must be addressed for URLLC. As an approach to address this problem, members of our research group reported on a low latency HARQ method that uses channel state information (CSI) prior to channel decoding [6]–[8]. This method mitigates the increased transmission latency by requesting early retransmission before the channel decoding process is completed based on the CSI obtained prior to channel decoding.

On the other hand, a non-orthogonal HARQ method in which the retransmission packet and the subsequent initial packet are non-orthogonally multiplexed in the same channel based on superposition coding is reported in [9]–[11]. This method reduces the bandwidth loss (throughput loss) associated with retransmission compared to that for the conventional orthogonal HARQ that allocates an exclusive channel (time-frequency slot) to the retransmission packet. Furthermore, assuming URLLC, non-orthogonal HARQ reduces the transmission latency since retransmission does not incur a transmission delay for the subsequent packet [10]–[12]. We note that applying non-orthogonal HARQ using superposition coding to the low latency HARQ method that uses CSI prior to channel decoding is investigated in [8].

In [13], [14], members of our research group have reported a highly-efficient and low-latency HARQ method built on non-orthogonal multiple access (NOMA) for downlink URLLC while minimizing the performance loss for coexisting services (use cases) such as eMBB. When non-orthogonal HARQ using superposition coding is applied to the cellular system using NOMA with a successive interference canceller (SIC) [15] as the multiple access scheme, retransmission packets can be non-orthogonally multiplexed with the packets of other terminals. This enables more effective use of non-orthogonal HARQ at the system level where the superposition coding of multiple packets is applied to the set of user terminals that have a good channel relationship in terms of the SIC process. The reported method uses channel-aware time/frequency-domain instan-

Manuscript received February 16, 2023.
Manuscript revised May 22, 2023.
Manuscript publicized July 24, 2023.

[†]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 Holdings Corporation, Yokohama-shi, 224-8539 Japan.

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

DOI: 10.1587/transcom.2023EBT0005

Copyright © 2023 The Institute of Electronics, Information and Communication Engineers

taneous resource allocation including scheduling that multiplexes the URLLC packet to the best-effort eMBB packet within the same time-frequency block using superposition coding. In scheduling, it is important to obtain multi-user diversity between terminals by taking into account the instantaneous channel conditions of all candidate user terminals in all services. To achieve this, the proposed resource allocation metric is a throughput-based one for multiple services, which is based on [16], with service-dependent appropriate weighting, which is based on [16], [17], so that a low PER and low transmission latency in URLLC can be explicitly guaranteed. Furthermore, in order to suppress the decoding error at the receiver caused by the error in transmission rate control due to the outdated CSI, i.e., channel quality indicator (CQI), service-dependent backoff operation in the transmission rate control was introduced. With the above configuration, the PER and transmission delay time of URLLC are improved while suppressing the deterioration in the system throughput of eMBB operated at the same time.

In this paper, we propose extending the NOMA-based highly-efficient low-latency HARQ method [13], [14] for URLLC to the case with inter-base station cooperation. In the proposed method, delay-sensitive URLLC packets are preferentially multiplexed with eMBB packets in the same channel using superposition coding to reduce the transmission latency of the URLLC packet while alleviating the throughput loss in eMBB. Although data transmission to a URLLC terminal is conducted by multiple base stations based on inter-base station cooperation, the proposed method allocates the radio resources (bandwidth and power) to URLLC terminals at each base station independently to achieve the short transmission latency required for URLLC. To avoid excessive radio resource assignment to URLLC terminals due to independent resource assignment at each base station, which may result in throughput degradation in eMBB terminals, we employ an adaptive path-loss-dependent weighting approach in the scheduling-metric calculation. This reduces the PER and transmission delay time of URLLC terminals by utilizing the effect of inter-base station coordination, while avoiding base stations far away from a URLLC terminal from allocating more radio resources than necessary to that URLLC terminal. Computer simulation results quantitatively show the performance gain when using the proposed method. We note that the proposed method can be operated in conjunction with the low-latency HARQ method in [6]–[8] using CSI before channel decoding. However, this paper assumes normal HARQ in which the retransmission request is sent after channel decoding is completed to assess the basic performance of the proposed method.

The remainder of this paper is organized as follows. First, Sect. 2 presents the proposed method. Section 3 shows the numerical results based on computer simulations. Finally, Sect. 4 concludes the paper.

2. Proposed Method

In this paper, we consider the downlink cellular system where eMBB and URLLC are operated simultaneously within the system coverage. The set of base stations in the system coverage is denoted as $\mathcal{B}$. The set of terminals that receive eMBB service is denoted as $\mathcal{K}_{\text{eMBB}}$. The set of terminals that receive URLLC service is denoted as $\mathcal{K}_{\text{URLLC}}$. The set of all terminals in the system coverage is denoted as $\mathcal{K} = \mathcal{K}_{\text{eMBB}} \cup \mathcal{K}_{\text{URLLC}}$. The set of downlink frequency blocks that is shared by eMBB and URLLC is denoted as $\mathcal{F}$.

It is assumed that the inter-base station cooperation allows for downlink data transmission to a terminal from multiple base stations. The set of base stations that transmit downlink data to a terminal, in other words, the set of terminals to which a base station transmits downlink data, is determined based on the terminal association method in [18]. First, each terminal selects the base station with the highest received signal power as its 'primary' base station. Inversely, the set of terminals having the same primary base station is denoted as the 'primary' terminal set for that base station. Term η $(0 < \eta \leq 1)$ is defined as the threshold value in terms of the instantaneous received signal power and is used to determine which subset of base stations is used for downlink cooperative data transmission for each user terminal. At each scheduling interval, each base station sets the association threshold value as η times the instantaneous channel gain of the terminal under the worst channel conditions in its primary terminal set. The candidate user terminals for downlink data transmission scheduling at each base station are set to the user terminals with instantaneous channel gain exceeding the threshold value in addition to the primary terminal set of that base station. The set of terminals that is subject to downlink data transmission at base station $b \in \mathcal{B}$ at time t is denoted as $\mathcal{K}_b(t)$. The set of cooperating base stations that transmit downlink data to terminal $k \in \mathcal{K}$ at time t is denoted as $\mathcal{B}_k(t)$. As η is set to be decreased, $|\mathcal{K}_b(t)|$ and $|\mathcal{B}_k(t)|$ are expected to increase, which means more extensive inter-base station cooperation.

Next, we explain the downlink data transmission method using multiple base stations in the proposed method. For eMBB terminals, as shown in Fig. 1(a), the information bits to be transmitted to each terminal are distributed among cooperating base stations and they are transmitted as independent coded packets (codewords) from each cooperating base station in order to maximize the inter-base station cooperation effect. To achieve such inter-base station cooperative data transmission, it is necessary to distribute the payload to be transmitted among multiple base stations in a coordinated manner and perform scheduling accordingly.

However, the process for distributing the payload among multiple base stations leads to an increase in transmission delay time. To avoid this for delay-sensitive URLLC, the proposed method assumes the downlink data transmission using multiple base stations for URLLC ter-

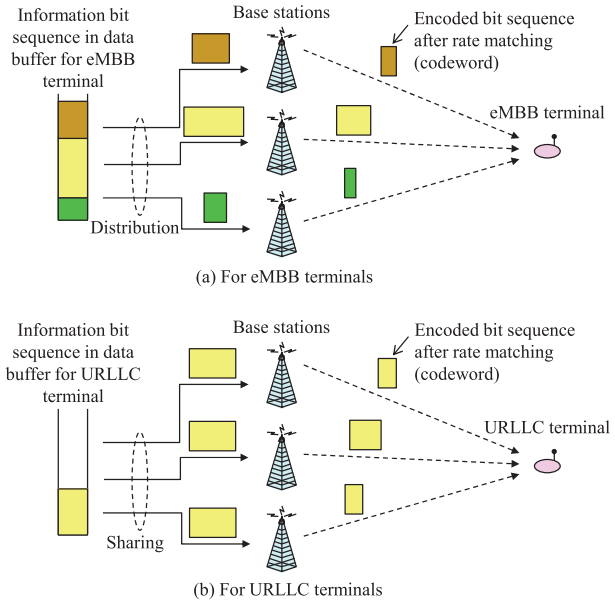

Fig. 1 Inter-base station cooperative data transmission method for eMBB and URLLC.

minals as shown in Fig. 1(b). The same information bit sequence to be transmitted to the URLLC terminal is shared by all the cooperating base stations. Each base station independently performs scheduling and transmits all or part of the encoded bit sequence (codeword) for the same information bit sequence using the allocated bandwidth through rate matching. This minimizes the transmission delay time associated with inter-base station cooperation. However, the independent scheduling at each base station may cause unnecessary radio resource use. The proposed method mitigates this problem by introducing an adaptive weighting factor in the scheduling metric, as described below.

By extending the scheduling metric for the case without inter-base station cooperation based on the weighted proportional fairness [13], [14] to the case with inter-base station cooperation, the scheduling metric of frequency block $f \in \mathcal{F}$ at time t for base station b is obtained as

$$\lambda_{b,f}\left(\mathcal{U}_{b,f}, \mathcal{P}_{b,f}; t\right) = \sum_{k \in \mathcal{K}_b(t)} \frac{\delta_{k,b}(t) r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}, \mathcal{P}_{b,f}; t\right)}{\left|\mathcal{K}_{i_k}\right| R_k(t-1)}. \tag{1}$$

The scheduling metric, $\lambda_{b,f}\left(\mathcal{U}_{b,f}, \mathcal{P}_{b,f}; t\right)$, is a function of allocated terminal set $\mathcal{U}_{b,f}$ and power allocation set $\mathcal{P}_{b,f}$, and the $\mathcal{U}_{b,f}$ and $\mathcal{P}_{b,f}$ that maximize it are selected. In this paper, power allocation $\mathcal{P}_{b,f}$ among non-orthogonal multiplexed terminals is determined for a given $\mathcal{U}_{b,f}$ based on the fixed power allocation method in [19] for the sake of simplicity, and frequency block f is assigned to terminal set $\mathcal{U}_{b,f}$ that maximizes the metric in (1). Term $\left|\mathcal{U}_{b,f}\right|$ is fixed at 1 in orthogonal multiple access (OMA) and $\left|\mathcal{U}_{b,f}\right|$ can be greater than 1 in NOMA. Here, $R_k(t)$ is the average throughput of terminal k at time t. Term $r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}, \mathcal{P}_{b,f}; t\right)$ is the transmission rate to terminal k at frequency f for base station b, which corresponds to the expectation of the instantaneous throughput (channel capacity) assuming NOMA with a SIC for the given set of scheduled terminals, $\mathcal{U}_{b,f}$, and that of allocated power levels, $\mathcal{P}_{b,f}$, for terminals in $\mathcal{U}_{b,f}$. Term i_k is the service indicator of user terminal k; thus, i_k = 'eMBB' when $k \in \mathcal{K}_{\mathrm{eMBB}}$, while i_k = 'URLLC' when $k \in \mathcal{K}_{\mathrm{URLLC}}$. Here, $\delta_{k,b}(t)$ is a weighting factor introduced with the intention of prioritizing frequency block allocation to URLLC terminals. In [13], [14], weighting factor $\delta_{k,b}(t)$ is determined only by the classification of URLLC and eMBB terminals. In this case, $\delta_{k,b}(t)$ is represented as

$$\delta_{k,b}(t) = \begin{cases} \delta_{\mathrm{URLLC}}, & i_k = \mathrm{URLLC} \\ \delta_{\mathrm{eMBB}}, & i_k = \mathrm{eMBB} \end{cases}. \tag{2}$$

By setting δ_{URLLC} greater than δ_{eMBB}, allocation to URLLC terminals can be prioritized. Without loss of generality, δ_{eMBB} is set to 1 in this paper. In the following, this method is referred to as the fixed δ method.

However, when the fixed δ method is applied to the low-latency inter-base station cooperation method shown in Fig. 1(b) for URLLC, there is a concern that a base station that is a long distance from a URLLC terminal will allocate an excessive amount of radio resources to that URLLC terminal due to independent scheduling at each base station. This deteriorates the eMBB terminal throughput. Therefore, the proposed method appropriately controls the weighting factor of the scheduling metric for URLLC terminals based on the path gain between each base station and URLLC terminal. This method is referred to as the adaptive δ method. In the proposed adaptive δ method, $\delta_{k,b}(t)$ for URLLC terminal k is represented as

$$\delta_{k,b}(t) = \max\left(\frac{G_{k,b}(t)}{\overline{G}_k(t)}, \delta_{\mathrm{eMBB}}\right). \tag{3}$$

Here, $G_{k,b}(t)$ is the path gain between URLLC terminal k and base station b at time t. Assuming that B_{avg} is defined as the number of base stations used to calculate the average path gain between URLLC terminal k and nearby base stations, $\overline{G}_k(t)$ is the average of B_{avg} path gains in increasing order between URLLC terminal k and the surrounding base stations. The adaptive δ method avoids base station b far from URLLC terminal k allocating more radio resources than necessary to that URLLC terminal by setting $\delta_{k,b}(t)$ to an appropriately low value, while maintaining the effect of a low PER and short transmission delay time for the URLLC terminal using the inter-base station cooperation. As a result, the throughput degradation in the eMBB terminal is reduced.

Next, we explain the transmission rate control for each of the URLLC and eMBB terminals in the proposed method. Assuming the application of a SIC, the maximum transmission rate that terminal k can correctly decode at frequency block f at time t from base station b for given $\mathcal{U}_{b,f}(t)$ and $\mathcal{P}_{b,f}(t)$, $r^{\mathrm{IDEAL}}_{k,b,f}\left(\mathcal{U}_{b,f}(t), \mathcal{P}_{b,f}(t); t\right)$, is represented as [13]–[15], [19]

$$r^{\mathrm{IDEAL}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right)=$$
$$\begin{cases} W\log_2\left(1+\dfrac{g_{k,b,f}(t)p_{k,b,f}(t)}{\sum\limits_{j\in\mathcal{U}_{b,f}(t),g_{k,b,f}(t)<g_{j,b,f}(t)} g_{k,b,f}(t)p_{j,b,f}(t)+1}\right), & k\in\mathcal{U}_{b,f}(t) \\ 0, & k\notin\mathcal{U}_{b,f}(t) \end{cases}. \tag{4}$$

Here, W is the bandwidth of the frequency block, $g_{k,b,f}(t)$ is the instantaneous channel gain of terminal k normalized by the noise power including the inter-cell interference power at frequency block f from base station b at time t, and $p_{k,b,f}(t)$ is the assigned transmission power.

However, the actual transmission rate, $r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right)$, is determined based on the CQI report, $\tilde{g}_{k,b,f}(t)$, which contains error due to measurement accuracy and CQI reporting delay time in a time-varying channel. Considering the initial transmission case, the instantaneous throughput of terminal k at time t is degraded due to the CQI error as follows.

In this paper, we assume that the coded bit sequence (codeword) for a given information bit sequence is transmitted across multiple allocated frequency blocks, as in 4G LTE/LTE-Advanced [1], [2], and 5G NR [3]. In this case, the maximum decodable rate is determined by the sum of Shannon's channel capacity, i.e., the maximum obtainable mutual information for a given channel, among all the allocated frequency blocks [20], [21]. First, we consider the eMBB terminal. For eMBB terminals, each base station transmits an independent codeword (coded data packet) as shown in Fig. 1(a). Therefore, instantaneous throughput $r_{k,b}(t)$ of terminal k obtained from base station b at time t is represented as

$$r_{k,b}(t)=$$
$$\begin{cases} \sum\limits_{f\in\mathcal{F}} r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right), & \sum\limits_{f\in\mathcal{F}} r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right) \\ & \le \sum\limits_{f\in\mathcal{F}} r^{\mathrm{IDEAL}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right) \\ 0, & \text{Otherwise} \end{cases}. \tag{5}$$

Thus, if the sum of the transmission rate, $r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right)$, among all frequency blocks is greater than that of the correctly-decodable maximum rate, $r^{\mathrm{IDEAL}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right)$, terminal k fails to decode the received packet and the throughput becomes zero. Term $r_k(t)$ is the sum of $r_{k,b}(t)$ for all cooperating base stations.

Next, we consider URLLC terminals. In this paper, we assume that fixed R_{URLLC} bits are transmitted to the URLLC terminal with a time interval of T_{URLLC}. The R_{URLLC} value is relatively small, and all R_{URLLC} bits are channel coded to generate codeword C_{URLLC}. Based on Fig. 1(b), each base station performs independent scheduling and rate matching operations on the same C_{URLLC}. The rate matching is actualized using puncturing or repetition of coded bits along with the appropriate choice for the modulation level, in 4G LTE/LTE-Advanced [1], [2], and 5G NR [3]. So, each base station transmits the independently-rate-matched C_{URLLC} using the allocated frequency blocks. Therefore, the maximum decodable rate is determined by the sum of the channel capacity among all the allocated frequency blocks of all the cooperating base stations. On the other hand, since codeword C_{URLLC} conveys R_{URLLC} information bits, the maximum decodable rate needs to be higher than $R_{\mathrm{URLLC}}/T_{\mathrm{p}}$ where T_{p} is the scheduling period that equals the packet length. Therefore, a URLLC packet with R_{URLLC} bits at time t is correctly decoded when the following condition is satisfied.

$$\begin{aligned} R_{\mathrm{URLLC}}/T_{\mathrm{p}} &\le \sum_{b\in\mathcal{B}_k(t)}\sum_{f\in\mathcal{F}} r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right) \\ &\le \sum_{b\in\mathcal{B}_k(t)}\sum_{f\in\mathcal{F}} r^{\mathrm{IDEAL}}_{k,b,f}\left(\mathcal{U}_{b,f}(t),\mathcal{P}_{b,f}(t);t\right). \end{aligned} \tag{6}$$

Here, $r_k(t)$ is $R_{\mathrm{URLLC}}/T_{\mathrm{p}}$ if the URLLC packet is correctly decoded, otherwise it is 0.

According to the above discussion, we consider reducing the probability of the decoding error by reducing the transmission rate in advance. Here, the effects of decoding errors can differ between eMBB, where decoding error is mainly observed as throughput loss, and URLLC, where decoding error severely affects the transmission delay time and PER. Therefore, in the proposed method, the transmission rate, $r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f},\mathcal{P}_{b,f};t\right)$, is defined as

$$r^{\mathrm{TX}}_{k,b,f}\left(\mathcal{U}_{b,f},\mathcal{P}_{b,f};t\right)=$$
$$\begin{cases} W\log_2\left(1+\dfrac{\alpha_{i_k}\tilde{g}_{k,b,f}(t)p_{k,b,f}(t)}{\sum\limits_{j\in\mathcal{U}_{b,f}(t),\tilde{g}_{k,b,f}(t)<\tilde{g}_{j,b,f}(t)} \tilde{g}_{k,b,f}(t)p_{j,b,f}(t)+1}\right), & k\in\mathcal{U}_{b,f}(t) \\ 0, & k\notin\mathcal{U}_{b,f}(t) \end{cases}. \tag{7}$$

By multiplying the predicted value of the signal-to-noise plus interference ratio (SINR) after SIC processing based on CQI $\tilde{g}_{k,b,f}(t)$ by a service-specific constant, α_{eMBB} or α_{URLLC} where $0\le\alpha_{\mathrm{eMBB}},\alpha_{\mathrm{URLLC}}\le 1$, the transmission rate is intentionally reduced. As a result, the occurrence of packet decoding errors due to CQI error caused by, e.g., the time fluctuation in the channel can be suppressed. So, the increase in the transmission delay time due to retransmission occurrence can also be reduced. The proposed method assumes that α_{URLLC} is set lower than α_{eMBB}. This suppresses the occurrence of URLLC packet decoding errors during initial transmission, and aims to achieve both a low PER and low transmission latency at a high level. In the proposed method, URLLC-prioritized resource allocation is performed using NOMA, and eMBB packets can be multiplexed within the same channel by superposition coding. Therefore, the adverse effects on the bandwidth allocation to eMBB due to such a safe-side transmission rate control in URLLC can be

alleviated. The use of small α_{URLLC} may also indirectly contribute to alleviating unnecessary radio resource allocation at far away base stations in inter-base station cooperation using independent scheduling at each base station shown in Fig. 1(b).

The operation at retransmission in the proposed method is described below. Suppose that a decoding error occurs in the initial transmission packet from base station b at time t_0 of a certain terminal k. Packet retransmission is performed at $t > t_0$, as a result of the resource allocation process for a retransmitted packet for terminal k based on the metric in (1). The resource allocation process for the retransmitted packet is basically the same as that for the initial packet. However, the transmission rate at retransmission is kept the same as the initial transmission rate. At the terminal receiver, the packet combining based on incremental redundancy is performed. As a result, the decodable transmission rate becomes a sum of $r_{k,b,f}^{\mathrm{IDEAL}}\left(\mathcal{U}_{b,f}(t), \mathcal{P}_{b,f}(t); t\right)$ for the initial transmission and each retransmission. For URLLC terminals, when T_{URLLC} elapses after the occurrence of the R_{URLLC}-bit payload, that payload is discarded and the transmission of the next R_{URLLC}-bit payload starts.

3. Numerical Results

The system-level performance of the proposed method is evaluated based on computer simulations. Table 1 gives the major simulation parameters. We assume a 36-MHz system bandwidth with 50 frequency blocks ($|\mathcal{F}| = 50$) with bandwidth W of 720 kHz. Universal frequency reuse is assumed among cells. We assume a heterogeneous networks [22] in which low-transmission-power pico base stations are overlaid onto the coverage area of a high-transmission-power macro base station. We assume that there are eMBB and URLLC terminals that are multiplexed within the system bandwidth. The macro and pico base stations, and terminals of each service are placed in random locations within a wrap-around 2×2-square kilometer system coverage area based on the Poisson point process (PPP). The node densities for the macro and pico base stations, and terminals receiving eMBB and URLLC are set to 1, 4, 40, and 20 per square kilometer, respectively. The transmission power level of macro and pico base stations is 46 and 30 dBm, respectively. As the propagation model, distance dependent path loss with the decay factor of 3.76 assuming the carrier frequency of 2 GHz [23]; lognormally distributed random shadowing with the standard deviation of 8 dB and inter-site correlation of 0.5; and 6-path Rayleigh fading with the maximum Doppler frequency of 5.55 Hz and the rms delay spread of 1 μs are assumed. The receiver noise power density of the terminal including a noise figure is set to −165 dBm/Hz.

Table 1 Simulation parameters.

System bandwidth		36 MHz
Number of frequency blocks		50
Coexisting services (use cases)		eMBB and URLLC
Node density	Macro base station	1 / km^2
	Pico base station	4 / km^2
	eMBB terminal	40 / km^2
	URLLC terminal	20 / km^2
Transmission power	Macro base station	46 dBm
	Pico base station	30 dBm
Distance-dependent path loss (including antenna gain)		114.1+37.6log$_{10}(r)$, r: kilometers
Shadowing		Lognormal 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		−165 dBm/Hz
Traffic model	eMBB terminal	Full buffer
	URLLC terminal	T_{URLLC} = 10 ms, R_{URLLC} = 1,000 bits
Scheduling interval		0.25 ms
Minimum interval of retransmission in HARQ		1 ms
Delay in CQI report		10 ms
Maximum number of non-orthogonally multiplexed terminals		2 per frequency block

The resource allocation (scheduling and power allocation) interval, which equals the packet length, is set to 0.25 ms. The maximum number of non-orthogonally multiplexed terminals per frequency block is set to two in the proposed method, which should be sufficient to obtain the most from the potential gain of NOMA based on [15] and [19]. We note that the proposed method includes OMA in addition to NOMA, since if $|\mathcal{U}_f(t)|$ is set to one based on the evaluation results of the proposed scheduling metric in (1), this means that frequency block f is operated by OMA at time t. When NOMA is applied, the fixed power allocation method [19] is used and 80% of the transmission power is allocated to terminals experiencing poor channel conditions. The remaining 20% of the power is allocated to terminals experiencing good channel conditions. The minimum retransmission interval is 1 ms [3]. The reporting delay of CQI, which is used for transmission rate control, is set to 10 ms.

The traffic model of the eMBB terminal is a full buffer, and the 100-ms average terminal throughput is measured. The system throughput defined by the geometric mean terminal throughput, which corresponds to the proportional fair criterion, is used as the main performance measure of eMBB. On the other hand, we assume that fixed $R_{\mathrm{URLLC}} = 1,000$ information bits are generated every $T_{\mathrm{URLLC}} = 10$ ms to respective URLLC terminals. The information bits that have not been transmitted correctly are discarded 10 ms after the occurrence. The main performance measures for URLLC are the transmission delay time for the correct decoding of a 1,000-bit URLLC payload, and the outage probability of the correct decoding of the 1,000-bit URLLC payload during the required transmission delay time of 1 ms. Parameters δ_{eMBB} and α_{eMBB} in the proposed method are fixed at 1.0 and 0.7, respectively. The $\delta_{k,b}(t)$ of the URLLC terminal are evaluated by comparing the fixed δ method using (2) with $\delta_{\mathrm{URLLC}} = 30$ and the proposed adaptive δ method with $B_{\mathrm{avg}} = 10$. Term α_{URLLC} is parameterized. Parameter η, which determines the range of cooperation between base stations, is set to 0.1. For comparison to the proposed method, we also evaluate the conventional method without inter-base station cooperation and the case where

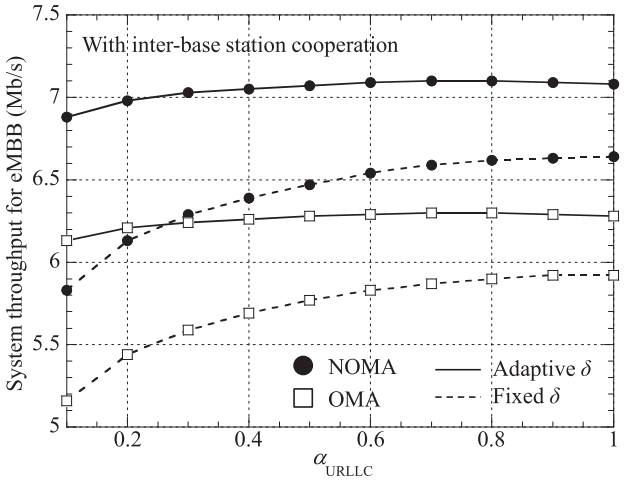

Fig. 2 System throughput for eMBB as a function of α_{URLLC}.

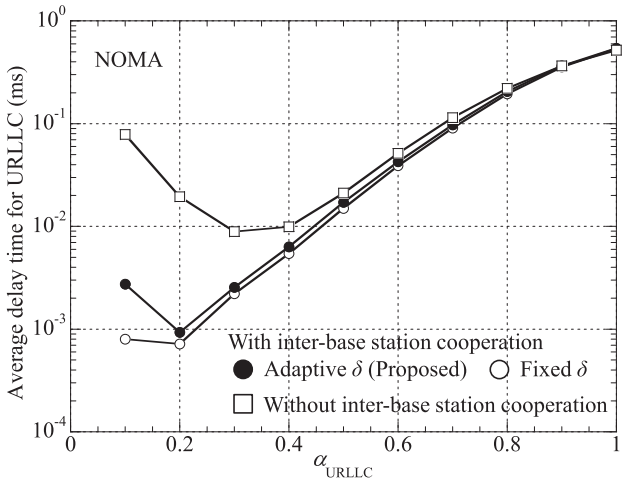

Fig. 3 Average delay time for URLLC as a function of α_{URLLC}.

only OMA is used in the proposed method with the inter-base station cooperation.

Figure 2 shows the system throughput for eMBB as a function of α_{URLLC}. Inter-base station cooperation is assumed and the fixed and adaptive δ methods for the NOMA and OMA cases, respectively, are tested. Overall, the system throughput for eMBB is increased when using NOMA compared to OMA. This is because the effect of NOMA, which multiplexes multiple terminals within the same bandwidth, is effective in increasing the system throughput, defined as the geometric mean terminal throughput, which requires fairness among eMBB terminals, while achieving prioritized radio resource allocation to URLLC terminals. Reducing α_{URLLC} is effective in increasing the correct decoding probability for URLLC data. However, it leads to increased radio resource allocation to URLLC terminals, which degrades the system throughput for eMBB. This is because reducing the transmission rate means using a combination of a lower coding rate and lower-order modulation (the use of QPSK instead of 16QAM for example) in an actual system, which increases the number of complex symbols required to transmit a given information bit sequence. The degradation in the eMBB system throughput due to the reduction of α_{URLLC} is significant for the fixed δ method. On the other hand, the proposed adaptive δ method alleviates the degradation in the eMBB system throughput when reducing α_{URLLC}. This is because the fixed δ method results in inefficient use of radio resources, especially at base stations far away from the destination URLLC terminal, since all cooperating base stations equally perform preferential scheduling for the URLLC terminal. On the other hand, the adaptive δ method controls weight δ of the scheduling metric that prioritizes URLLC terminals according to the path gain with each base station appropriately and this avoids the excessive allocation of radio resources at base stations experiencing poor channel conditions with the target URLLC terminal. As a result, the adaptive δ method suppresses the reduction in bandwidth allocated to eMBB terminals at the system level.

Figure 3 shows the average delay time for URLLC as a function of α_{URLLC}. The transmission delay time is defined as the time required to complete correct decoding for all 1,000-bit URLLC payloads during each 10-ms time interval. The average transmission delay time is calculated for the case where 1,000 information bits could be transmitted correctly within 10 ms. In this figure, NOMA is assumed and we evaluate the cases with and without inter-base station cooperation for both using the fixed or adaptive δ methods. The increase in the transmission delay time when using an α_{URLLC} value greater than 0.2 is due to frequent packet retransmissions. This is a result of excessive transmission rate settings due to CQI errors. The inter-base station cooperation significantly reduces the average transmission delay time for URLLC especially when using a low α_{URLLC} value. The increase in the transmission delay time for URLLC when inter-base station cooperation is not applied is mainly due to the insufficient system bandwidth to convey the 1,000-bit URLLC payloads to the URLLC terminal under very poor channel conditions. The maximum amount of radio resources allocated per URLLC terminal can be increased using inter-base station cooperation, and this reduces the transmission delay time for URLLC terminals located in such poor environments. With inter-base station cooperation, the adaptive δ method avoids the excessive allocation of radio resources at base stations experiencing poor channel conditions with the target URLLC terminal, compared to that for the fixed δ method. However, this reduced radio resource allocation to URLLC terminals in the adaptive δ method does not significantly degrade the average transmission delay time for URLLC as shown in Fig. 3. This is because the adaptive δ method reduces the unnecessary radio resource allocation to URLLC terminals by appropriately controlling weight δ of the scheduling metric that prioritizes URLLC terminals according to the path gain with each base station. As a result of reduced radio resource allocation to URLLC terminals while maintaining approximately the same average transmission delay time for URLLC, the proposed adaptive δ method increases the sys-

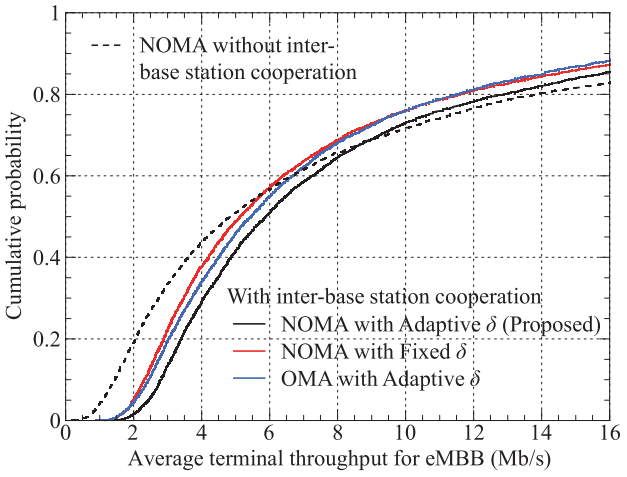

Fig. 4 Distribution of average terminal throughput for eMBB.

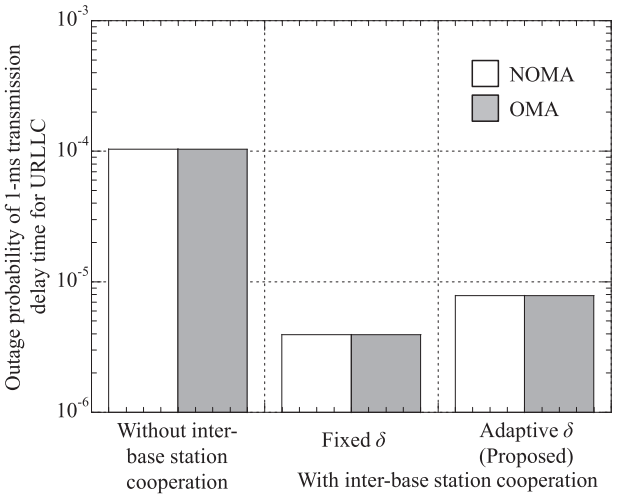

Fig. 5 Outage probability of transmission delay time of 1 ms for URLLC.

tem throughput for eMBB as already shown in Fig. 2. Based on Fig. 3, in the following evaluations, α_{URLLC} is set to 0.2 and 0.3 for the cases with and without inter-base station cooperation, respectively.

Figure 4 shows the cumulative distribution of the average throughput for eMBB terminals. The purpose of this evaluation is to clarify fully how the throughput of individual terminals experiencing different channel conditions changes in the proposed method compared to that in the conventional method, which cannot be fully understood from only the system throughput shown in Fig. 2. From Fig. 4, the proposed NOMA with inter-base station cooperation using the adaptive δ method achieves the highest average throughput of eMBB terminals over almost the entire cumulative probability range compared to that for the fixed δ method, OMA, and the case without inter-base station cooperation. The improvement in performance from the case without inter-base station cooperation is especially significant in the low cumulative probability region. This indicates that the proposed inter-base station cooperation method is especially effective in improving the fairness of throughput among terminals.

Figure 5 shows the outage probability that a 1,000-bit URLLC payload transmission cannot be completed within the required transmission delay time of 1 ms. The proposed method significantly reduces the outage probability compared to the case without inter-base station cooperation, and achieves an outage probability of approximately 10^{-5}, which is almost equivalent to that for the fixed δ method. Meanwhile, the proposed method with adaptive δ achieves a much higher eMBB system throughput compared to the case with the fixed δ method as shown in Fig. 2. We note that the outage probability levels are approximately the same between the proposed NOMA and OMA in the case with inter-base station cooperation. The first reason for this is that we set the proposed parameters in NOMA such as δ and α_{URLLC} so that the transmission delay time for URLLC is reduced as much as possible. In principle, the transmission performance of URLLC can be degraded with NOMA compared to OMA due to the inter-terminal interference with non-orthogonal terminal multiplexing. However, this impact is considered small due to the following reason. The outage probability for the transmission delay time of 1 ms for URLLC is determined by the user terminals experiencing poor channel conditions. Such user terminals are allocated high transmission power in NOMA, while the interference from non-orthogonally multiplexed packets of user terminals near the base station cannot be removed by the SIC (this is because the decoding order of the SIC is based on the order of user terminals experiencing poor channel conditions). However, for user terminals experiencing very poor channel conditions, the received signal power from the set of serving base stations is low, while the inter-cell interference power is very high, so the impact of interference power from the serving base stations due to non-orthogonal packet multiplexing in NOMA is very low. Based on the reasons above, the outage probability for the transmission delay time of 1 ms for URLLC when the proposed NOMA is used is almost the same as that for OMA.

4. Conclusion

In this paper, we extended our previously reported highly-efficient and low-latency HARQ method, which non-orthogonally multiplexes the initial and retransmission packets of URLLC terminals to packets of eMBB terminals based on the downlink NOMA, to inter-base station cooperation systems. Although data transmission to URLLC terminals is conducted by multiple base stations based on inter-base station cooperation, the proposed method allocates the radio resources to URLLC terminals at each base station independently to achieve the short transmission latency required for URLLC. To avoid excessive radio resource assignment to URLLC terminals due to independent resource assignment at a base station that is far from the destination URLLC terminal, we employ the adaptive path-loss-

dependent weighting approach in calculating the scheduling metric. This achieves appropriate radio resource assignment to URLLC terminals while reducing the PER and transmission delay time thanks to the inter-base station cooperation. The reduced resource usage to the URLLC terminals results in improvement in the throughput for eMBB terminals. Computer simulations revealed quantitatively that the outage probability of a 1-ms transmission delay time for URLLC is small enough for the proposed method, while increasing the system throughput for eMBB compared to the case with fixed weight coefficients and to the case without using inter-base station cooperation.

References

[1] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband, 2nd ed., Academic Press, 2008.
[2] E. Dahlman, S. Parkvall, and J. Sköld, 4G: LTE/LTE-Advanced for Mobile Broadband, 2nd ed., Academic Press, 2013.
[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] D.N. Rowitch and L.B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes," IEEE Trans. Commun., vol.48, no.6, pp.948–959, June 2000.
[6] Y. Imamura, D. Muramatsu, Y. Kishiyama, and K. Higuchi, "Low latency hybrid ARQ method using channel state information before channel decoding," Proc. APCC2017, Perth, Australia, Dec. 2017.
[7] K. Taniyama, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission prior to channel decoding with multistage decision," Proc. IEEE VTC2019-Fall, Honolulu, U.S.A., Sept. 2019.
[8] K. Miura, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission before channel decoding based on superposition coding," Proc. ISPACS 2019, Taipei, Taiwan, Dec. 2019.
[9] R. Zhang and L. Hanzo, "Superposition-coding aided multiplexed hybrid HARQ scheme for improved link-layer transmission efficiency," Proc. IEEE ICC2009, Dresden, Germany, June 2009.
[10] F. Takahashi and K. Higuchi, "HARQ for predetermined-rate multicast channel," Proc. IEEE VTC 2010-Spring, Taipei, Taiwan, May 2010.
[11] Y. Hasegawa and K. Higuchi, "Bi-directional signal detection and decoding for hybrid ARQ using superposition coding," Proc. IEEE VTC2011-Fall, San Francisco, U.S.A., Sept. 2011.
[12] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Non-orthogonal HARQ for URLLC: Design and analysis," IEEE Internet Things J., vol.8, no.24, pp.17596–17610, 2021. doi: 10.1109/JIOT.2021.3081698.
[13] R. Kobayashi, Y. Yuda, and K. Higuchi, "NOMA-based highly-efficient low-latency HARQ method for URLLC," Proc. IEEE VTC2021-Fall, Virtual conference, Sept.–Oct. 2021.
[14] R. Kobayashi, Y. Yuda, and K. Higuchi, "Highly-efficient low-latency HARQ built on NOMA for URLLC: Radio resource allocation and transmission rate control aspects," IEICE Trans. Commun., vol.E106-B, no.11, pp.–, Nov. 2023 (Advance publication).
[15] 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.
[16] T. Shikuma, Y. Yuda, and K. Higuchi, "NOMA-based optimal multiplexing for multiple downlink service channels to maximize integrated system throughput," IEICE Trans. Commun., vol.E103-B, no.11, pp.1367–1374, Nov. 2020.
[17] C. Wengerter, J. Ohlhorst, and A.G.E. von Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," Proc. IEEE VTC2005-Spring, pp.1903–1907, Stockholm, Sweden, May–June 2005.
[18] T. Shikuma, Y. Yuda, and K. Higuchi, "NOMA-based inter-base station cooperative scheduling method among multiple service channels to maximize integrated system throughput," Proc. IEEE VTC2020-Spring, Virtual conference, May 2020.
[19] 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.
[20] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.
[21] L. Wan, S. Tsai, and M. Almgren, "A fading-insensitive performance metric for a unified link quality model," Proc. IEEE WCNC 2006, Las Vegas, USA, April, 2006.
[22] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," IEEE Wireless Commun. Mag., vol.18, no.3, pp.10–21, June 2011.
[23] 3GPP, TR 25.814 (V7.0.0), "Physical layer aspects for Evolved UTRA," June 2006.

Ryota Kobayashi received the B.E. and M.E. degrees from Tokyo University of Science, Noda, Japan in 2021 and 2023, respectively. In 2023, he joined NTT DOCOMO, INC.

Takanori Hara received the B.E., M.E., and Ph.D. degrees in engineering from The University of Electro-Communications, Tokyo, Japan, in 2017, 2019, and 2022, respectively. Since April 2022, he has been with the Department of Electrical Engineering, at Tokyo University of Science, Chiba, Japan, where he is currently an Assistant Professor. His current research interests are grant-free access, compressed sensing, and MIMO technologies.

Yasuaki Yuda received the B.E. and M.E. degrees from Tokyo University of Science, Japan in 1997 and 1999, respectively. He received a Ph.D. from Tokai University, Japan, in 2014. Since 1999, he has been with the Matsushita Electric Industrial Co., Ltd., Panasonic Corporation and Panasonic Holdings 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, 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 member of the IEEE.