Latency of Uplink Non-orthogonal Multiple Access Grant-Free Transmission with Multiuser Detection in Base Station

Show Shiono*, Yukitoshi Sanada*, Ryota Kimura†, Hiroki Matsuda†, Ryo Sawai†
*Dept. of Electronics and Electrical Engineering,
Keio University, Yokohama, Japan
Email: s_shiono@snd.elec.keio.ac.jp, sanada@elec.keio.ac.jp
†Fundamental Technology Research and Development Division1, R&D Center,
Sony Corporation, Tokyo, 141-8610 Japan
Email: {Ryota.Kimura, Hiroki.Matsuda, Ryo.Sawai}@sony.com

***Abstract*—In this paper, the latency of non-orthogonal multiple access (NOMA) grant-free (GF) transmission in uplink is evaluated by link and system level simulations. In order to cancel interference among NOMA signals, multiuser detection (MUD) consisting of minimum mean square error demodulation followed by successive interference cancellation is applied. Retransmission happens only when a packet error occurs after serial interference cancellation (SIC). Therefore, the latency is reduced as compared to GF transmission without MUD. Numerical results obtained through system level simulation show that the application of MUD in uplink GF transmission improves the latency and increases the successful transmission rate without retransmission by a factor of about 3.0.**

I. Introduction

As smartphones and tablet devices penetraite in our daily life and the reserch and the development of ultra low latency applications such as self-driving are actively carried out, more strict requirements to mobile communications are imposed in recent years. The fifth generation mobile communication (5G) is designed to satisfy these requirements and its services are going to be launched in 2020 [1]. One of issues under discussion for satisfying low latency conditions is grant-free (GF) transmission [2]. In the GF transmission a user equipment (UE) transmits a packet without a grant from a base station (BS). In this way, a packet can be transmitted with lower latency than grant-based (GB) transmission. However, multiple UEs may transimit simultaneously in a same resource block. Therefore, the reliability and latency evaluation of the GF transmission have been investigated these days [3]–[7].

On the other hand, non-orthogonal multiple access (NOMA) in 5G has been paid large attention [8]–[10]. In the NOMA the signals of multiple UEs are transmitted at the same resource block and a receiver extracts its desired signals by using a multiuser detection technique. Recent studies suggest that the system capacity of the NOMA can be lager as compared to that of orthogonal multiple access (OMA) [11], [12].

Recent literature have investigated NOMA GF transmission with the use of interference cancellation (IC) [8], [13]. However, few of them conduct system level capacity evaluation [14]–[16], and none of them investigates transmission in conjunction with IC. Therefore, this paper evaluates the latency of NOMA GF transmission when multiuser detection (MUD) is applied in an uplink. This research conducts link level simulation to obtain the block error rate (BLER) curves of NOMA transmission and then evaluates the latency of NOMA GF transmission based on the obtained BLER curves.

This paper is organized as follows. Section 2 describes the system model. Section 3 explains the simulation conditions. Numerical results obtained through computer simulation are then presented. Section 4 gives our conclusions.

II. System Description

A. Link Level Model

1) Uplink Transmission: An uplink transmission model shown in Fig. 1 is assumed. U UEs are served by a BS. Here, the number of transmit antennas of each UE and the number of receive antennas of the BS are both set to one. The turbo code standardized for long term evolution (LTE) is used as a forward error correction code [17]. The turbo encoder standardized for LTE is composed of one interleaver and two 8-state encoders that are connected in parallel. For the input of N_s bits, the outputs of the turbo encoder consist of three length-N_s streams called "Systematic", "Parity 1", "Parity 2", and 12 tailbits by trellis termination processing. The bits of each stream encoded in the turbo encoder are interleaved in the sub-block interleaver. After interleaving, the output bits of the "Systematic" stream are followed by the output bits of the "Parity 1" stream and the "Parity 2" stream that are arranged alternately bit by bit, and the encoded bits are series of length-$(3Ns+12)$ bits. Each UE constructs a 2^M QAM symbol with M bits, therefore, the length of a packet is $D = \frac{3N_s+12}{M}$ symbols. After applying inverse discrete Fourier transform (IDFT) to a transmission symbol sequence, a cyclic prefix (CP) consisting of the last N_g samples of the symbol sequence is inserted as a guard interval at the beginning of the symbol sequence. It is then transmitted from a UE.

In the BS, the CP is removed from the received signal and it is put into a discrete Fourier transform (DFT) block with the size of N_c. The received signal on the k-th subcarrier $\{Y[k]; 0 \leq k \leq N_c - 1\}$ is written as

$$Y[k] = \sum_{u=1}^{U} H_u[k]S_u[k] + N[k] \tag{1}$$

where u $(1 \leq u \leq U)$ is the UE index, $S_u[k]$ is the transmit symbol of the u-th UE on the k-th subcarrier, $H_u[k]$ is the channel response between the u-th UE and the BS on the k-th subcarrier, and $N[k]$ is the additive white Gaussian noise (AWGN) with a mean of 0 and a variance of σ^2 on the k-th subcarrier.

The frame format in GF transmission is presented in Fig. 2. It is assumed here that the number of resource blocks allocated for GF transmission is G_f, the number of OFDM symbols per

978-1-7281-1220-6/19/$31.00 ©2019 IEEE

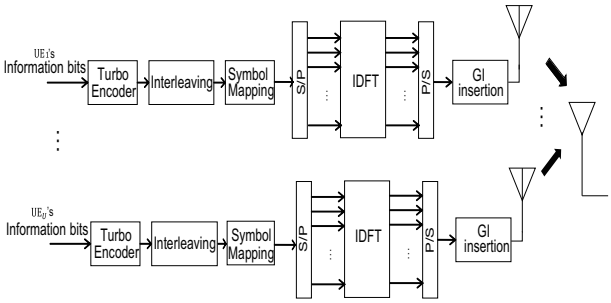

Fig. 1. Uplink transmission.

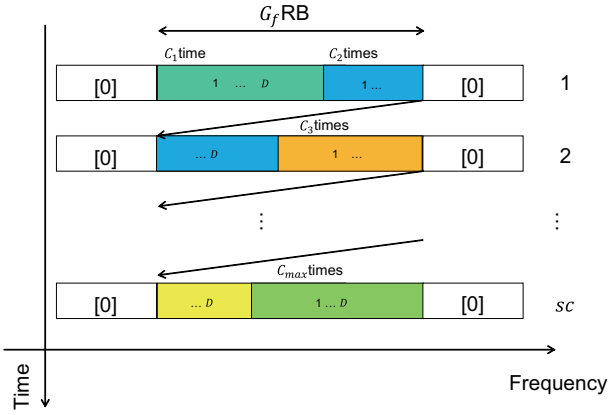

Fig. 2. Frame format.

frame is $Sc = 7$, and the number of subcarriers per resource block is 12. The d-th symbol is transmitted repeatedly C_{max} times, where d ($1 \leq d \leq D$) is the index of the transmitted symbol in one packet. Therefore, the subcarrier index, λ_{dc}, for the c-th repetition of the d-th symbol is given as

$$\lambda_{dc} = (cD + d) \pmod{12G_f}, \tag{2}$$

and the number of repetition, C_{max}, is then given as

$$C_{max} = \left\lfloor \frac{84G_f}{D} \right\rfloor, \tag{3}$$

where $\lfloor * \rfloor$ is the floor function and represents the largest integer that is less than $*$.

2) MMSE Demodulation: A MUD model is shown in Fig. 3. As the same symbol is transmitted C_{max} times, minimum mean square error (MMSE) demodulation can be applied over those symbols. The signals from all the UEs pass through independent fading channels. At the receiver, MMSE demodulation is first applied to the received signals [18].

$$\hat{\mathbf{S}}_d^{(0)} = \mathbf{W}_d \mathbf{Y}_d \tag{4}$$

where

$$\mathbf{W}_d = (\mathbf{H}_d^H \mathbf{H}_d + \sigma^2 \mathbf{I})^{-1} \mathbf{H}_d^H, \tag{5}$$

$\hat{\mathbf{S}}_d^{(0)}$ is the d-th transmit symbol vector, $\mathbf{Y}_d$ is the d-th received signal vector, $\mathbf{H}_d$ is the channel matrix corresponding to the d-th transmit symbol, and $\mathbf{W}_d$ is the weight matrix for the d-th transmit symbol. Also, $[\ ]^H$ is the Hermitian matrix and $\mathbf{I}$ is the unit matrix with the size of $U \times U$. The elements of these matrices and vectors are given as follows.

$$\mathbf{Y}_d = [Y[\lambda_{d1}]\ Y[\lambda_{d2}]\ \ldots\ Y[\lambda_{dC_{max}}]]^T, \tag{6}$$

$$\mathbf{H}_d = [\mathbf{H}_{d1}\ \mathbf{H}_{d2}\ \ldots\ \mathbf{H}_{dU}], \tag{7}$$

$$\hat{\mathbf{S}}_d^{(0)} = [\hat{S}_{d1}^{(0)}\ \hat{S}_{d2}^{(0)}\ \ldots\ \hat{S}_{dU}^{(0)}]^T. \tag{8}$$

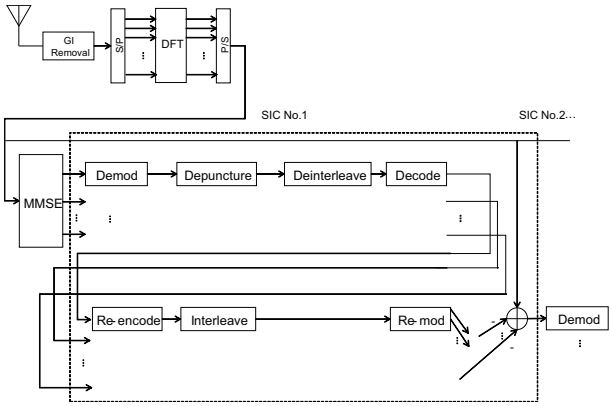

Fig. 3. Multiuser detection model.

Here, $[\ ]^T$ represents the transposed matrix, $Y[\lambda_{dc}]$ is the received signal at the c-th repetition of the d-th transmit symbol, and $\hat{S}_{du}^{(0)}$ is the d-th transmit symbol of the u-th UE obtained by MMSE demodulation. Furthermore, $\mathbf{H}_{du}$ is the channel response vector of the u-th UE for the d-th transmit symbol and is written as

$$\mathbf{H}_{du} = [H_u[\lambda_{d1}]\ H_u[\lambda_{d2}]\ \ldots\ H_u[\lambda_{dC_{max}}]]^T, \tag{9}$$

where $H_u[\lambda_{dc}]$ is the channel response from the u-th UE to the BS at the c-th repetition of the d-th transmit symbol.

The likelihoods of the d-th transmit symbol of the u-th UE is obtained as follows.

$$\Delta_{du1}^{MMSE} = \sum_{\hat{S}_{du} \in \{S_u\}|_{b_m^u=1}} \exp(-\tfrac{1}{\sigma^2} \|S_{du}^{(0)} - \hat{S}_{du}\|^2), \tag{10}$$

$$\Delta_{du0}^{MMSE} = \sum_{\hat{S}_{du} \in \{S_u\}|_{b_m^u=0}} \exp(-\tfrac{1}{\sigma^2} \|S_{du}^{(0)} - \hat{S}_{du}\|^2), \tag{11}$$

where Δ_{du1}^{MMSE} and Δ_{du0}^{MMSE} are the sums of the likelihood values for the m-th bit being "1" and "0" in the d-th symbol of the u-th UE, $\hat{S}_{du}$ is the symbol candidate of the u-th UE, and $\{S_u\}|_{b_m^u=1}$ and $\{S_u\}|_{b_m^u=0}$ are the sets of 2^M QAM symbols of which the m-th bit is equal to "1" and "0" in the signal of the u-th UE, respectively. Using Eqs. (10) and (11), a log-likelihood ratio (LLR) is calculated as follows.

$$L(b_m^u|\mathbf{Y}_d) = \log \frac{\Delta_{du1}^{MMSE}}{\Delta_{du0}^{MMSE}} \tag{12}$$

where $L(b_m^u|\mathbf{Y}_d)$ is the LLR for the m-th bit of the d-th symbol transmitted from the u-th UE. The LLRs for all the transmit symbols are input to the decoder and the signal of the u-th UE is then decoded.

3) Codeword Level SIC: The decoded signal is then re-encoded and the replica signals for all the UEs are generated for codeword level signal interference cancellation (SIC). The received signal of the u-th UE after SIC is written as

$$\mathbf{Y}_{du}^{(i)} = \mathbf{Y}_d - \mathbf{H}_{d\backslash u} \hat{\mathbf{S}}_{d\backslash u}^{(i-1)} \tag{13}$$

where $\mathbf{Y}_{du}^{(i)}$ is the d-th received signal vector of the u-th UE after the i-th SIC, $\mathbf{H}_{d\backslash u}$ is the channel matrix corresponding to the d-th transmit symbol except that of the u-th UE, and $\hat{\mathbf{S}}_{d\backslash u}^{(i)}$ is the demodulated symbol vector excluding the symbol

of the u-th UE after the i-th SIC for the d-th symbol. These vectors are given as follows.

$$\mathbf{Y}_{du}^{(i)} = [Y_u^{(i)}[\lambda_{d1}]\ Y_u^{(i)}[\lambda_{d2}]\ \dots\ Y_u^{(i)}[\lambda_{dC_{max}}]]^T, \quad (14)$$

$$\mathbf{H}_{d\setminus u} = [\mathbf{H}_{d1}\ \mathbf{H}_{d2}\ \dots\ \mathbf{H}_{d(u-1)}\ \mathbf{H}_{d(u+1)}\ \dots\ \mathbf{H}_{dU}]^T, \quad (15)$$

$$\hat{\mathbf{S}}_{d\setminus u}^{(i)} = [\hat{S}_{d1}^{(i)}\ \hat{S}_{d2}^{(i)}\ \dots\ \hat{S}_{d(u-1)}^{(i)}\ \hat{S}_{d(u+1)}^{(i)}\ \dots\ \hat{S}_{dU}^{(0)}]^T, \quad (16)$$

where $Y_u^{(i)}[\lambda_{dc}]$ is the received signal at the c-th repetition of the d-th transmit symbol of the u-th UE after the i-th SIC and $\hat{S}_{du}^{(i)}$ is the replica of the d-th transmit symbol of the u-th UE after the i-th SIC. The LLR is calculated for the received signal, $Y_u^{(i)}[\lambda_{dc}]$, after the application of the SIC and the likelihood values for the d-th transmit symbol of the u-th UE is obtained as follows.

$$\Delta_{du1}^{(i)} = \sum_{\hat{S}_{du}\in\{S_u\}|_{b_m^u=1}} \exp\left\{-\frac{1}{\sigma^2}\left(\sum_{C=1}^{C_{max}}\left|Y_u^{(i)}[\lambda_{dc}]-H_u[\lambda_{dc}]\hat{S}_{du}\right|^2\right)\right\}, \quad (17)$$

$$\Delta_{du0}^{(i)} = \sum_{\hat{S}_{du}\in\{S_u\}|_{b_m^u=0}} \exp\left\{-\frac{1}{\sigma^2}\left(\sum_{C=1}^{C_{max}}\left|Y_u^{(i)}[\lambda_{dc}]-H_u[\lambda_{dc}]\hat{S}_{du}\right|^2\right)\right\}, \quad (18)$$

where $\Delta_{du0}^{(i)}$ and $\Delta_{du0}^{(i)}$ are the sums of the likelihood values for the m-th bit being "1" and "0" in the d-th symbol of the u-th UE after the i-th SIC, $\hat{S}_{du}$ is the symbol candidate for the d-th symbol of the u-th UE, and $\{S_u\}|_{b_m^u=1}$ and $\{S_u\}|_{b_m^u=0}$ are the set of 2^M QAM symbols of which the m-th bit is equal to "1" and "0" in the signal of the u-th UE, respectively. The LLR is obtained as follows.

$$L(b_m^u|\mathbf{Y}_{du}^{(i)}) = \log\frac{\Delta_{du1}^{(i)}}{\Delta_{du0}^{(i)}} \quad (19)$$

where $L(b_m^u|\mathbf{Y}_{du}^{(i)})$ is the LLR for the m-th bit of the d-th transmit symbol of the u-th UE after the i-th SIC. The LLRs are input to the decoder and the transmit signal of the u-th UE is decoded.

In this paper, the application of the SIC is repeated several times to improve the accuracy of interference cancellation. Through the above link level simulation, the block error rate (BLER) curve on a fading channel with regards to a signal-to-noise ratio (SNR) is obtained.

B. System Level Model

1) NOMA GF Transmission: System level simulation evaluates the latency performance of NOMA GF transmission and GB transmission in the uplink. The system level packet transmission model is shown in Fig. 4. The frequency resources are divided into GF and GB slots in the frequency domain. UEs first transmit in the GF slot. If the transmitted packet is successfully demodulated at the BS, the UE can transmit a next packet in the GF slot at the next frame. However, since multiple UEs share the same slot in GF transmission, packet collision may occur. In this case, the signal-to-interference-and-noise ratio (SINR) of the packet is calculated and a packet error probability is derived using the BLER curve obtained through link level simulation. Based on the packet error probability, the packet reception is judged as an error.

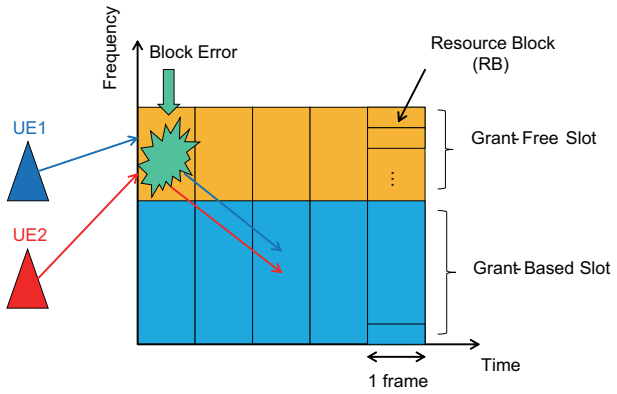

Fig. 4. Packet transmission model.

The UE then retransmit the same packet in the GB slot after receiving grant from the BS. Even though there is no packet collision, it retransmits the packet if the packet reception is failed in the GF slot. The GB transmission takes additional two slots for packet transmission.

2) Re-transmission Model: In the GB slot, resource blocks (RBs) are assigned to a UE with a packet to be retransmitted based on proportional fairness (PF) metrics. The PF metric of the u-th UE in the l-th RB of the f-th frame is given as

$$Q_u^{fl} = \Sigma_{k\in\{k^{fl}\}} C_{M_u}(E_u^k)/\bar{Q}_u, \quad (20)$$

where k^{fl} is the subcarrier index in the l-th RB of the f-th frame, C_M is the constellation constraint capacity (CCC) for the modulation order of M, M_u is the modulation order of the u-th UE, E_u^k is the SINR of the u-th UE in the k-th subcarrier, and $\bar{Q}_u$ is the average metric of the u-th UE. The CCC for the modulation order of M is given as [19], [20]

$$C_M(E) = M - \mathrm{E}_Y\left[\frac{1}{2^M}\Sigma_{m=1}^{M}\Sigma_{b=0}^{1}\Sigma_{S\in\{S_m^b\}}\log 2\left\{\frac{\Sigma_{\hat{S}\in\{S\}}\exp(-|Y-\sqrt{E}(\hat{S}-S)|^2}{\Sigma_{\hat{S}\in\{S_m^b\}}\exp(-|Y-\sqrt{E}(\hat{S}-S)|^2}\right\}\right], \quad (21)$$

where Y follows a normal distribution with a mean of zero and a variance of one, S corresponds to 2^M QAM constellation points, and S_m^b corresponds to constellation points when the m-th bit of M bits that is assigned to the 2^M QAM symbol is b.

A BLER curve on an AWGN channel with each modulation and coding scheme (MCS) index are evaluated beforehand and the minimum SINR required for RB allocation is determined. The coding efficiency is calculated from the MCS and the correction coefficient, β, is determined from the MCS [21]. In this research ARQ retransmission is assumed and the throughput per subcarrier is obtained by the following equation using the correction coefficient, β, determined from the MCS.

$$I_u = \Sigma_f\Sigma_l\Sigma_{k\in\{k_u^{fl}\}} C_{M_u}\left(E_u^k/\beta\right) \quad (22)$$

where $\{k_u^{fl}\}$ is a set of subcarrier indices in the l-th RB of the f-th frame to which the u-th UE is allocated and E_u^k is the SINR of the u-th UE on the k-th subcarrier. The effective SNR is obtained from the throughput from Eq. (22) as follows.

$$\gamma_{eff} = \beta I^{-1}(I_u) \quad (23)$$

The BLER on the AWGN channel is calculated by the effective SNR and the MCS. The packet is assumed to be retransmitted

with the probability of the BLER for the effective SNR and it is received with the probability of 1-BLER.

III. Numerical Results

A. Link Level Simulation

1) Simulation Conditions: The BLER versus the SNR in the GF transmission with multiuser detection is evaluated by link level simulation. The simulation conditions are shown in Table I. The number of data bits per packet is 80 bits. A turbo code is assumed for the error correction code [17]. The code rate is 1/3, the interleaver size is 80 bits, and the modulation scheme is QPSK. The channel bandwidth is 2.5 MHz and the subcarrier spacing is 15 kHz. The guard interval for OFDM symbols is 5.21μs for the first symbol and 4.69μs for or the following six symbols. The number of RBs allocated for GF transmission is assumed to be 12 RBs and the same encoded symbol is transmitted eight times in one frame over 144 subcarriers. On the receiving side, the SIC is applied five times after the MMSE demodulation as multiuser detection. A Log-MAP algorithm is applied for decoding and the number of decoding iterations is eight [22]. A six-path Rayleigh fading channel with a uniform delay profile is applied and channel estimation at the receiver side is assumed to be ideal. The number of trials is set to be 800000. For comparison, the BLER for a single user is also evaluated.

TABLE I
Simulation Conditions :Link Level

Information Bits/Packet	$D = 80$ bits
Error Correction Code	Turbo Code
Interleaver Size	80 bits
Code Rate	1/3
Modulation Scheme	QPSK/OFDM
Channel Bandwidth	2.5 MHz
Subcarrier Spacing	15kHz
GF Resource Block Size	$G_f = 12$ RBs
GB Resource Block Size	38 RBs
Number of Subcarriers	144
Guard Interval Length	5.21μs (1st), 4.69μs (2nd-7th)
MUD Scheme	MMSE + SIC
Decoding Scheme	Log-Map Algorithm
Number of Decoding Iterations	8
Channel Model	Six-path Rayleigh Fading
Channel Estimation	Ideal
Number of Trials	800000

2) BLER Performances: The BLER curves obtained by link level simulation are shown in Fig. 5. The number of UEs is two and the signal power of UE 2 normalized by that of UE 1 is 0.9. The SNR in the x-axis of the figure is defined as the ratio between the signal power and the noise power for each UE. From this figure, at a BLER of 10^{-1}, the difference from the single user's performance is within 0.5 dB for both of the UEs. Therefore, the BLER curve for the single user is applied in the system level simulation.

B. System Level Simulation

1) Simulation Conditions: The simulation conditions of the system level simulation is shown in Table II. The cell layout is assumed to be 19-hexagonal-cell and five UEs exist in each sector. The distance between BS antennas is 500 meters and the minimum distance between the BS and the UE is 35 meters. The transmission power is set to be 23 dBm and the distance dependant path loss is given as 128.1+37.6log10(R) dB (R km). The BS antenna height is 25 meters and the UE antenna height is 1.5 meters. The receiver noise density is -174 dBm/Hz and the shadowing standard deviation is 8 dB. A six-path Rayleigh fading channel with a uniform delay profile is assumed. The PF scheduling is applied as the scheduling algorithm in the GB slots. The ARQ retransmission interval is five slots. The system bandwidth is 9 MHz, the RB bandwidth is 180 kHz, and the number of subcarriers per RB is 12. The number of GF UEs is five users. The GF resource size is 12 RBs while the GB resource size is 38 RBs. The number of information bits per packet is 80 bits and the modulation scheme is QPSK. The packet is generated with a Poisson distribution and its coefficient is set to 0.05 or 0.2. The number of user drops is 50, the number of trials per drop is 10, and the number of frames per trial is 100. For comparison to the MUD, link level simulation without MUD is carried out. In this case, packet collision leads to a packet error.

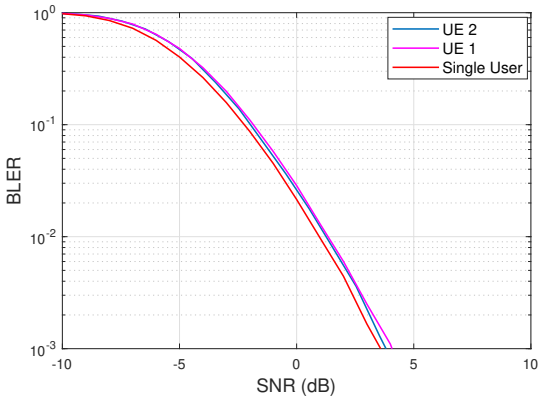

Fig. 5. BLER vs. SNR (2UEs, $P_2 = 0.9$).

2) Latency Evaluation: The latency performance obtained by system level simulation is presented in Figs. 6-7. In the figures, the y-axis indicates the cumulative probability and the x-axis is the number of frames required for successful transmission. When the Poisson coefficient is 0.05, the probability of packets with MUD that can be successfully transmitted in the GF slot is 0.05 larger. As the Poisson coefficient increases to 0.2, the number of generated packets per slot increases and the amount of improvement in the latency with MUD increases. When the Poisson coefficient is 0.20, the number of packets that can be successfully transmitted in the GF slot (i.e. at a latency of 1 slot) is about three times larger in the NOMA system.

IV. Conclusions

In this paper, the latency of the uplink NOMA GF transmission is evaluated. MUD consisting of the MMSE demodulation followed by the SIC is applied in the NOMA GF transmission and retransmission is carried out only when a packet error after the MUD occurs. In the link level simulation, the BLER performance with NOMA is obtained. The latency of the NOMA GF transmission is evaluated through system level simulation based on the BLER curve obtained through the link level simulation. Numerical results have shown that the application of MUD in the GF transmission reduces the latency. It has also been confirmed that the improvement in the packet latency is more significant when the number of generated transmission packets per slot is larger.

TABLE II
SIMULATION CONDITIONS :SYSTEM LEVEL

Cell Layout	19 hexagonal cell site
Inter-site Distance	500 m
Minimum Distance (BS - UE)	35 m
Distance Dependent Path Loss	128.1+37.6log10(R) dB (R km)
Maximum Transmit Power	23 dBm
BS height	25 m
UE height	1.5 m
Receiver Noise Density	-174 dBm/Hz
Shadowing Standard Deviation	8 dB
Channel Model	6 Path Rayleigh Fading
Scheduling Algorithm (GF slots)	PF Scheduling
Number of Users/Sector	5
System Bandwidth	9 MHz
RB Bandwidth	180 kHz
Number of Subcarrier per RB	12
GF Resource Block Size	12 RBs
GB Resource Block Size	38 RBs
Number of GF User	5
ARQ Retransmission Interval	5
Information Bits/Packet	80 bit
Modulation Scheme	QPSK
Poisson Coefficient (λ)	0.05, 0.2
User Drop	50
Trial/User Drop	10
Number of Frames/Trial	100

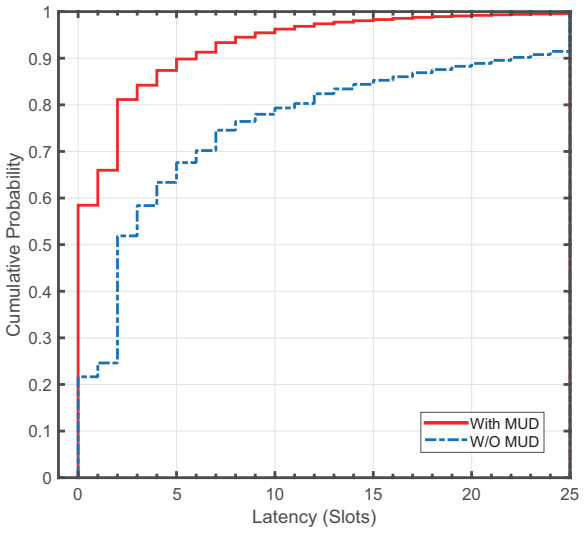

Fig. 6. CDF of latency (5 UEs, λ = 0.20).

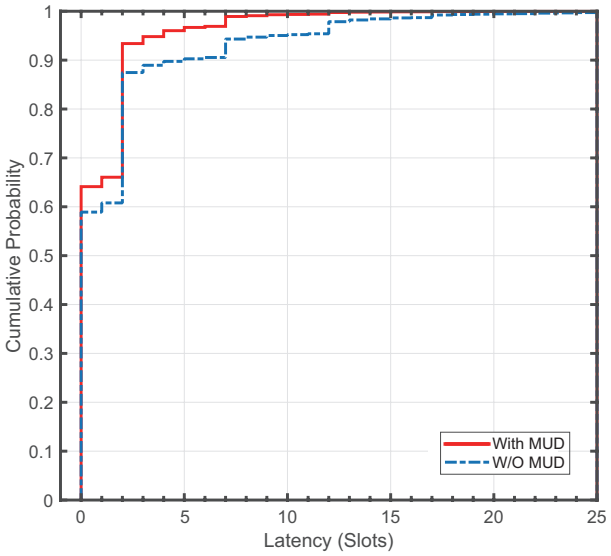

Fig. 7. CDF of latency (5 UEs, λ = 0.05).

REFERENCES

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

[2] "Technical Specification Group Radio Access Network; Study on Non-orthogonal Multiple Access (NOMA) for NR," 3GPP TS38.812, v16.0.0, Dec. 2018.

[3] L. G. Roberts, "ALOHA Packet System With and Without Slots and Capture," ACM SIGCOMM Comput. Commun. Rev., vol. 5, no. 2, pp. 28-42, April 1975.

[4] B. Singh, O. Tirkkonen, Z. Li and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," IEEE Wireless Commun. Lett. , vol. 7, no. 2, pp. 182-185, Apr. 2010.

[5] L. Dai, B. Wang, Y. Yuan, I. Han, S. Chih-Lin, 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, Sep. 2015.

[6] J. Zhang, L. Lu, Y. Sun, J. Liu, H. Yang, S. Xing, Y. Wu, J. Ma, I. B. F. Murias, and F. J. L. Hernando, "PoC of SCMA-Based Uplink Grant-Free Transmission in UCNC for 5G," IEEE Journal on Selected Areas in Commun. , vol. 35, no. 6, pp. 1353-1562, June 2017.

[7] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovác, and P. Mogensen, "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources," IEEE Access, vol. 6, pp. 23602-23611, April 2018.

[8] "Study on New Radio Access Technology - Physical Layer Aspects," 3GPP TS38.802 , v14.2.0, Sept. 2017.

[9] H. Osada, M. Inamori, and Y. Sanada, "Non-orthogonal Access Scheme over Multiple Channels with Iterative Interference Cancellation and Fractional Sampling in OFDM Receiver," IEICE Trans. on Communications, vol. E95-B, no. 11, pp. 3837-3844, Dec. 2012.

[10] K. Higuchi, A. Benjebbour, "Non-orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access," IEICE Trans. vol. E98-B, no. 3, pp. 403-414, March 2015.

[11] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "Systemlevel Performance Evaluation of Downlink Non-orthogonal Multiple Access (NOMA)," IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Sept. 2013.

[12] A. Benjebbour, Y. Kishiyama, and T. Nakamura, "Systemlevel Performance of Downlink NOMA for Future LTE Enhancements," IEEE Globeecom, Dec. 2013.

[13] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind Detection of SCMA for Uplink Grant-Free Multiple-Access," IEEE International Symposium on Wireless Communications Systems, Aug. 2014.

[14] M. Cheng, Y. Wu, Y. Li, Y. Chen, and L. Zhang, "PHY Abstraction and System Evaluation for SCMA with UL Grant-free Transmission," IEEE Vehicular Technology Conference, June 2017.

[15] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovács, and T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," IEEE Globecom, Dec. 2017.

[16] C. Boyd, R. Vehkalahti, O. Tirkkonen, "Grant-Free Access in URLLC with Combinatorial Codes and Interference Cancellation," IEEE Globecom 2018 Workshop, Dec. 2018.

[17] "LTE; Envolved Universal Terrestrial Radio Access; Multiplexing and channel coding," 3GPP TS36.212, v14.2.0, April 2017.

[18] Z. Xie, R. T. Short, and C. K. Rushforth, "A Family of Suboptimum Detectors for Coherent Multiuser Communications," IEEE Journal on Selected Areas in Comuunications, vol. 8, no. 4, pp. 683-690, May 1990.

[19] G. Gaire, G. Taricco, and E. Biglieri, "Capacity of bit-interleaved channels," Electronics Letters, vol. 32, no. 12, pp. 1060-1061, June 1996.

[20] X. L, Q. Fang, and L. Shi, "A Effective SINR Links to System Mapping Method for CQI Feedback in TD-LTE," IEEE 2nd International Conference on Computing, Control and Industrial Engineering, Aug. 2011.

[21] "LTE; Envolved Universal Terrestrial Radio Access; Physical layer procedures," 3GPP TS36.213, v9.2.0, June 2010.

[22] P. Robertson, E. Villebrum, and P. Hoehes, "A Comparison of Optimal and Sub-Optimal MAP Decoding Algorithms Operating in the Log Domain," IEEE International Conference on Commun. , vol. 2, pp. 1009-1013, June 1995.