Grant-free Multiple Access for Ultra-Reliable Low-Latency Communications in a Large-Scale Antenna System

Jong Hyun Kim, Kyung Jun Choi, Kwanghoon Lee, and Kwang Soon Kim
Department of Electrical and Electronic Engineering Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, Korea.
Email: {jhkim, kjchoi, ghl1016}@dcl.yonsei.ac.kr, ks.kim@yonsei.ac.kr

***Abstract*—In this paper, a grant-free multiple access scheme is proposed for ultra-reliable low-latency communications (URLLC) in a large-scale antenna system. The proposed scheme is designed for URLLC with sporadic uplink traffic characteristics and a statistically delay-optimal resource allocation is performed at the time of each user's service negotiation. According to the resource allocation, the frame configuration is adapted and a unique pilot pattern is allocated to each user for a grant-free multiple access. The performance of the proposed scheme is evaluated and compared with a conventional scheme and it shows that the proposed scheme can enhance the spectral efficiency while keeping the reliability and latency requirements compared to the conventional scheme.**

I. Introduction

Although the fourth generation (4G) cellular system has been successful in satisfying the demand for mobile broadband services, the fifth generation (5G) cellular system is envisioned to meet the exponentially growing demands for various mobile services in everytime and everywhere [1], in which the expected services are categorized as follows: enhanced mobile broadband (eMBB) services, massive machine-type communications (mMTC) services, and ultra-reliable and low latency communications (URLLC) services [2]. Among these services, the URLLC services has drawn much attention recently because it can provide a real-time interaction among machines and humans so that future services, such as augmented reality, industry automation, automated driving, tactile internet, real-time closed loop automated optimization and control among machines, and so on, can be realized in near future [3][4]. However, in order to realize such URLLC services, there exist technical challenges for designing an efficient multiple access scheme because typical data traffics need to be delivered with a very low latency (e.g., within a few hundred μs) and ultra-high reliability (e.g., frame error rate of 10^{-9}), which has not been even considered so far [5].

In the 4G cellular system, a grant-based multiple access with a user selection and rate adaptation based on the channel quality information (CQI) feedback is employed for uplink, by which the spectral efficiency can be significantly improved, and the spectral efficiency and the reliability can be simultaneously achieved by employing the hybrid automatic repeat and request (H-ARQ) [6]. However, these approaches are no longer valid options due to the low latency requirement of URLLC and a grant-free uplink multiple access scheme (without any scheduling based on instantaneous CQI feedback) is required. In addition, it needs to provide the ultra-high reliability without any retransmission while improving the average spectral efficiency compared to that in the 4G system for better cost-effective service penetration. A natural approach for achieving the ultra-reliability is to utilize diversity and a large diversity order is necessary to reduce fading margin into an acceptable level [7]. However, a simple use of diversity deteriorates the efficiency of a multiple access scheme and a carefully designed resource allocation scheme is required. On the other hand, a spectrally efficient grant-free multiple access scheme has been proposed in [8]. However, it is aimed to provide mMTC services so that the reliability resorts on H-ARQ and cannot achieve the low latency and ultra reliability requirements of URLLC simultaneously. Thus, it is highly desired to investigate a novel multiple access scheme that can provide high spectral efficiency while satisfying the low latency and ultra-high reliability requirements simultaneously.

Recently, a large-scale antenna system (LSAS) has drawn much attention, in which very large number of antennas are equipped at a base station (BS) to serve many users simultaneously and reliably [9] and such an LSAS can be an enabler for spectrally-efficient URLLC since it can provide a large diversity order as well as a large spectral efficiency simultaneously. In this paper, a grant-free multiple access scheme is proposed for URLLC in an LSAS. The proposed scheme is designed for URLLC with sporadic uplink traffic characteristics and a statistically delay-optimal resource allocation is performed at the time of each user's service negotiation. According to the resource allocation, the frame configuration is adapted and a unique pilot pattern is allocated to each user for a grant-free multiple access. Performance evaluation of the proposed scheme compared with conventional schemes are also provided for typical URLLC service scenarios in various channel environments.

II. System model

In Table. I, typical use cases for URLLC services in [3] are summarized, in which the traffic characteristics, such as the packet size and packet arrival rate and model, and the requirements, such as the latency and reliability, are found

978-1-5090-1325-8/16/$31.00 ©2016 IEEE

TABLE I
USE CASES IN THE 5GPPP WHITE PAPERS

White paper	Use case	Data rate	Latency	Reliability
Automotive	Automated Overtake	Up to 1Mbps	10ms	10^{-5}
	Cooperative Collision	Up to 1Mbps	Trajectory handshake: 100ms Status updates: 10ms	10^{-5}
	High Density Platooning	Up to 1Mbps	10ms	10^{-5}
	See-Through	10Mbps	50ms	10^{-2}
	Bird's Eye View	40Mbps	50ms	10^{-2}
Factories of the Future	Time-critical optimization	Low to high	Ultra-low	Ultra-high
	Remote control	Low to high	Less critical	High
	Connected Goods	Low	Less critical	Low
e-Health	Remote surgery	Up to 100Mbps	5ms	10^{-4}

to be very diverse. For example, the automotive use case requires a medium data rate (1Mbps) with a very low tolerance on errors (10^{-5}) and latency ($10ms$) for the critical driving situation while it requires high data rate (up to 40Mbps) with a medium tolerance on errors (10^{-2}) and latency (50ms) for multimedia video streamings.

Among the above service and traffic examples, sporadically arriving large-size packets are the most challenging ones because the randomness in currently accessing users is very high but high spectral efficiency with ultra-high reliability needs to be provided in order to deliver large-size packets within the latency requirements. Thus, such sporadic large-size packets are assumed for designing the proposed multiple access scheme in this paper.

In order to provide URLLC service, an uplink LSAS consisting of a BS with M antennas, and U single-antenna users is considered as illustrated in Fig. 1. The users want the quality of service (QoS) on their own sporadic data traffic (rate, latency, and reliability). In order to modeling the sporadic data traffic, we assume that the packet arrival of each user follows an independent Poisson arrival with average arrival rate of P_j. Also, each arrived packet needs to be delivered to the BS within the latency requirement reliably (satisfying the outage constant).

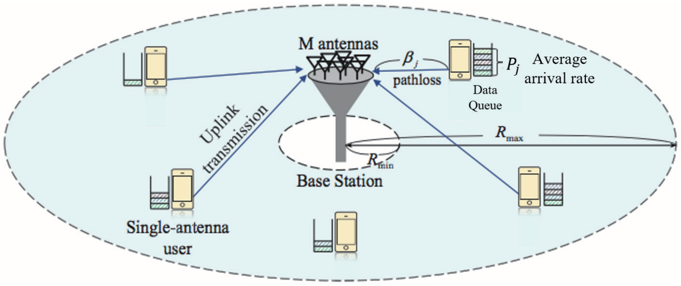

Fig. 1. System model

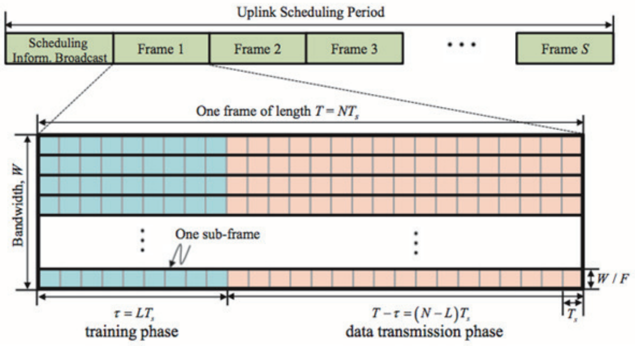

Fig. 2. Frame model

A two-phase frame structure with training and data transmission phases is illustrated in Fig. 2. The frame of time length T seconds and bandwidth W Hz is divided into F equal-bandwidth sub-frames by partitioning frequency domain by using the orthogonal division multiplex access (OFDMA), single-carrier frequency domain multiple access (SC-FDMA) or any good one of the newly considered waveforms [10]. The sub-frame consists of N symbols of time period T_s seconds. In the training phase of time period $\tau = LT_s$ seconds, an awaken user j sends L unique dedicated training symbols with power of p_j^{tr}, and the BS estimates the uplink channel. Then, in the data transmission phase of the remaining time period $T - \tau = (N - L)T_s$ seconds, the awaken user j transmits $N - L$ data symbols to the BS with power of p_j^{dt} in a space division multiple access (SDMA) manner.

Contrary to a granted transmission, our transmission strategy does not include any grant information to the users. Instead, users with packets to be delivered directly try to access the uplink at the earliest allocated resource, regardless of the channel state and the states of other users. This is why we call it a grant-free transmission. Our grant-free transmission has two-folded advantages over its counterpart: 1) it can minimize the latency by removing uplink-downlink switching for a grant, and 2) it can provide reliability to some degree if the static scheduling policy is well designed. So, it is suitable for applications requiring ultra-latency and ultra-reliability (urgent messages, connected cars, telemedicine, etc).

A. Training Phase

Suppose that U users havr their dedicated pilot sequences of length L. This allocation is performed at the beginning of a cell association or service negotiation step. Suppose that $\overline{\mathcal{S}}[f;t]$ denotes the set of allocated users in sub-frame f of frame t and $\mathcal{S}[f;t] \subset \overline{\mathcal{S}}[f;t]$ denotes the set of actually transmitting users. Note that at most M users are assumed to transmit.[1] Assume that $|\mathcal{S}[f;t]| = K$ in the sequel. For equalizing the difference of all users' channel estimation quality (maximizing the worst), the *channel-inversely power-controlled pilots* are assumed similarly as in [11], in which the average received signal energy of the users in $\mathcal{S}[f;t]$ is set to the common target received energy, $\overline{p^{\mathrm{tr}}}$. So, the transmit energy at the training phase is set by $p_j^{\mathrm{tr}} = \beta_j^{-1}\overline{p^{\mathrm{tr}}}$.

It is known that the Welch bound equality (WBE) sequence can provide the minimum sum cross-correlation property, and we assume that a sequence set satisfying the WEB is adopted for the pilots. Then, the mean square error (MSE) of the minimum mean square error (MSE) channel estimator is given by [11]

$$\sigma_{\mathrm{tr}}^2 = \frac{1+(K-L)^+\overline{p^{\mathrm{tr}}}}{1+\max\{L,K\}\overline{p^{\mathrm{tr}}}}, \tag{1}$$

where $(x)^+ = \max\{x, 0\}$.

B. Data Transmission Phase

After the training phase, the information-bearing signal is transmitted and it is detected by using a simple maximum ratio combining (MRC) or zero-forcing (ZF) receiver. By treating the estimated CSI as if it were the true CSI, the MRC or ZF receiver are given by respectively

$$\mathbf{F} = \begin{cases} \widehat{\mathbf{G}}_{\mathcal{S}[f;t]}\left(\widehat{\mathbf{G}}_{\mathcal{S}[f;t]}^H\widehat{\mathbf{G}}_{\mathcal{S}[f;t]}\right)^{-1}, & \text{if ZF,} \\ \widehat{\mathbf{G}}_{\mathcal{S}[f;t]}, & \text{if MRC,} \end{cases} \tag{2}$$

where $\widehat{\mathbf{G}}_{\mathcal{S}[f;t]} \in \mathcal{C}^{K\times M}$ is the estimated channel between K users and M antennas at the BS. Treating the interference as Gaussian random variables, the achievable rate of user k (bits/symbol) during sub-frame f of frame t is given by

$$R_k[f;t] \approx \log_2\left(1+\gamma_k[f;t]\chi_k\right), \tag{3}$$

where $\chi_k = \mathbb{1}(k \in \mathcal{S}[f;t])$ denotes the random variable indicating whether user k actually transmits when associated in the sub-frame f at time t and $\gamma_k[f;t]$ is given as

$$\gamma_k[f;t] = \\ \begin{cases} \frac{L(M-K)p_k^{\mathrm{dt}}\beta_k\overline{p^{\mathrm{tr}}}}{\left(1+(K-L)^+\overline{p^{\mathrm{tr}}}\right)\left(1+\sum\limits_{j\in\mathcal{S}[f;t]}p_j^{\mathrm{dt}}\beta_j\right)+L\overline{p^{\mathrm{tr}}}}, & \text{if ZF,} \\ \frac{L(M-1)p_k^{\mathrm{dt}}\beta_k\overline{p^{\mathrm{tr}}}}{\left(1+\max\{L,K\}\overline{p^{\mathrm{tr}}}\right)\left(1+\sum\limits_{j\in\mathcal{S}[f;t]}p_j^{\mathrm{dt}}\beta_j\right)-L\overline{p^{\mathrm{tr}}}p_k^{\mathrm{dt}}\beta_k}, & \text{if MRC.} \end{cases} \tag{4}$$

Algorithm 1: Optimal Scheduling Policy

```
Input: ${\{\beta_j\}}_{j=1}^U$
Output: $\mathcal{O}^\star, \mathcal{D}^\star$
Sort $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_U$.
– First Part: Find candidate scheduling groups
Set $q \leftarrow 1$.
for $1 \leq q_1 + q_2 \leq U,\ 0 \leq q_2 - q_1 \leq M-1$ do
    $\mathcal{O}_q \leftarrow \{q_1, q_1+1, \ldots, q_1+q_2-1\}$.
    Calculate $\Omega_q(\mathcal{O}_q)$ by using (7).
    $q \leftarrow q+1$.
end
– Second Part: Solve the binary integer programming
Construct $\mathbf{c}$ and $\mathbf{S}$ with (9) and (10).
Solve the LP (8) with relaxing $\mathbf{x} \in [0,1]^{C\times 1}$ and let $\mathbf{x}^\star$
    be its optimal solution.
Find the index set $\mathcal{Q} = \{q | [\mathbf{x}^\star]_q = 1\}$.
Compute
    $$D_q = \frac{\Omega_q^{-1}(\mathcal{O}_q)}{\sum_{i\in\mathcal{Q}}\Omega_i^{-1}(\mathcal{O}_i)},\ \forall q \in \mathcal{Q},$$
Return $\mathcal{O}^\star \leftarrow \{\mathcal{O}_q\}_{q\in\mathcal{Q}}$, and $\mathcal{D}^\star \leftarrow \{D_q\}_{q\in\mathcal{Q}}$.
```

III. Statistically delay-optimal resource allocation

A. Grant-Free Scheduling Policy

A grant-free scheduling policy for S frames is defined as $(\mathcal{O}, \mathcal{D})$, where $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_Q\}$ denotes the set of scheduling groups and $\mathcal{D} = \{D_1, D_2, \ldots, D_Q\}$ denotes the set of scheduling portions. Due to the nature of the grant-free uplink scheduling, the output of the grant-free scheduling policy should satisfy the following two constraints: 1) $\mathcal{O}_p \bigcap \mathcal{O}_q = \emptyset$ if $\forall p \neq q$, and 2) $\bigcup_{q=1}^Q \mathcal{O}_q = \{1, 2, ..., U\}$. Also, the scheduling portion is defined as

$$D_q = \frac{1}{FS}\sum_{t=1}^S\sum_{f=1}^F \mathbb{1}\{\overline{\mathcal{S}}[f;t] = \mathcal{O}_q\}, \tag{5}$$

so that it must satisfy $\sum_{q=1}^Q D_q = 1$.

Note that D_q is the portion of the sub-frames allocated to scheduling group q, $\mathcal{O}_q$, during one scheduling period consisting of FS sub-frames. In fact, since D_qFS is not an integer, $\lfloor D_qFS \rfloor$ sub-frames may be allocated. We assume that SF is sufficiently large so that the error $\lfloor D_qFS \rfloor - D_qFS$ is negligibly small.

B. Optimal Algorithm

In practical scenarios, the packet size, which needs to be delivered to a BS within one scheduling period, i.e., the network latency S, is determined by each user's service and an admission control is required when each service is negotiated. However, for simplicity, we assume that each user's packet size is determined to meet the reliability for a given scheduling policy and maximize the spectral efficiency, i.e., the optimization problem can be written as

[1] If more than M users transmit, it is treated as an outage event.

$$
\begin{aligned}
\text{(P)}\ \ &\min_{\mathcal{O},\mathcal{D}}\ T_j, \quad \text{subject to} \\
&\Pr\left((N-L)\sum_{t=1}^{S}\sum_{f=1}^{F} R_j[f;t] \geq \mathcal{T}_j\right) \geq 1-\epsilon, \forall j, \\
&\mathcal{O}_p \cap \mathcal{O}_q = \emptyset,\ \bigcup\nolimits_{p=1}^{Q} \mathcal{O}_p = \mathcal{U}, \\
&\sum\nolimits_{q=1}^{Q} D_q = 1,\ D_q \geq 0, \forall q,
\end{aligned}
$$

where T_j denotes the packet size for user j.

Suppose that $\mathcal{O} = \{\mathcal{O}_1, ..., \mathcal{O}_Q\}$ is given. From the reliability constraint, we have

$$
\begin{aligned}
&\Pr\left(\frac{N-L}{WTS}\sum_{t=1}^{S}\sum_{f=1}^{F} R_j[f;t] \geq \frac{\mathcal{T}_j}{WTS}\right) \\
&\qquad\qquad = \Pr\left(T_j \leq \frac{WTS\left(1-\frac{L}{N}\right)}{\eta} D_q \overline{R}_j\right),
\end{aligned}
$$

where $\overline{R}_j = \frac{\sum_{t=1}^{S}\sum_{f=1}^{F} R_j[f;t]}{\sum_{t=1}^{S}\sum_{f=1}^{F} 1_{j\in S[f;t]}}$ and $\eta = WT_s/F \geq 1$ denotes the bandwidth inefficiency (such as the cyclic prefix overhead). Note that $\overline{R}_j$ is a random variable whose cumulative distribution function $F_j(\cdot)$ is determined by the given scheduling policy, the long-term CSIs, and the power profile $\{(p_j^{\mathrm{tr}}, p_j^{\mathrm{dt}})|j = 1, ..., U\}$. Thus, for each scheduling policy, the allowed packet size T_j of user j can be evaluated as

$$
T_j = \frac{WTS(1-L/N)}{\eta} F_j^{-1}(\epsilon). \tag{6}
$$

Define

$$
\begin{aligned}
\Omega_q(\mathcal{O}_q) &= \min_{j\in\mathcal{O}_q} T_j \\
&= \frac{WTS(1-L/N)}{\eta} D_q \min_{j\in\mathcal{O}_q} F_j^{-1}(\epsilon).
\end{aligned} \tag{7}
$$

Here, the power profile within the subchannel for each user, i.e., $(p_j^{\mathrm{tr}}, p_j^{\mathrm{dt}})$ can be determined to maximize (7) similarly as in [11]. Also since it is proportional to the scheduling portion D_q, it is straightforward that the optimal scheduling policy is determined as $D_q = \Omega_q^{-1}(\mathcal{O}_q)/\sum_{q'=1}^{Q}\Omega_{q'}^{-1}(\mathcal{O}_{q'})$ and $T_{\min} = \min_j T_j$ is given as

$$
T_{\min} = \frac{WTS(1-L/N)}{\eta} \frac{1}{\sum_{q'=1}^{Q}\Omega_{q'}^{-1}(\mathcal{O}_{q'})}.
$$

Finally, the optimization problem is equivalent to

$$
\text{P-eq}\quad \underset{\mathcal{O}=\{\mathcal{O}_1,...,\mathcal{O}_Q\}}{\text{minimize}}\ \sum\nolimits_{q=1}^{Q} \frac{1}{\Omega_q(\mathcal{O}_q)}.
$$

Fortunately, this optimization problem (P-eq) has a similar form of (34) in [11] so that we can directly utilize this technique. For the completeness, we describe it briefly. First, the optimization problem (P-eq) can be transformed into a binary integer programming (BIP) with the following generic form:

$$
\underset{\mathbf{x}}{\text{minimize}}\ \ J(\mathbf{x}) = \mathbf{c}^T\mathbf{x}, \tag{8a}
$$

$$
\text{subject to}\ \ \mathbf{S}\mathbf{x} = \mathbf{1}_{C\times 1},\ \mathbf{x} \in \{0,1\}^{C\times 1}, \tag{8b}
$$

where $\mathbf{S} = [s_{uq}]$ is the $U \times C$ state matrix,

$$
s_{uq} = \begin{cases} 1, & u \in \mathcal{O}_q, \\ 0, & \text{otherwise}, \end{cases} \tag{9}
$$

$\mathbf{c}$ is the $C \times 1$ cost vector given by

$$
\mathbf{c} = \left[\frac{1}{\Omega_1(\mathcal{O}_1)}, \frac{1}{\Omega_2(\mathcal{O}_2)}, \cdots, \frac{1}{\Omega_C(\mathcal{O}_C)}\right]^T, \tag{10}
$$

where $\mathbf{1}_{C\times 1}$ is the $C \times 1$ all-one vector, $C = |\{\mathcal{O}_q\,|1 \leq |\mathcal{O}_q| \leq M, \mathcal{O}_q \subset \mathcal{U}\}|$ denotes the number of candidate scheduling groups. The optimizing variable $\mathbf{x}$ informs which candidate scheduling groups are selected, i.e., if $x_q = 1$, the corresponding candidate scheduling group $\mathcal{O}_q$ is selected as one of the optimal scheduling groups. Such a BIP has been widely researched in literature and a variety of efficient algorithms are summarized in [12]. Unfortunately, finding the optimal solution in a BIP is known as NP-hard in general. However, due to the special structure of our BIP (consecutive ones property), it will be shown that a linear programming (LP) relaxation using $\mathbf{x} \in [0,1]^{C\times 1}$ does not affect the optimality. So, the optimal solution of (x) is obtained the LP relaxation. The optimal algorithm is summarized at Algorithm 1. Finally, we can find the optimal training length L by trying the above algorithm for candidate values, similarly as in [11].

IV. Performance evaluation

In this section, we present some numerical results to verify the superiority of the proposed uplink scheduling policy. One frame is set to occupy $10MHz$ and $1ms$ in the frequency and time domains and consists of $F = 80$ sub-frames with $125KHz$ and $1ms$. The number of symbols in each sub-frame is set to $N = 100$ by assuming $= 1.25$ (25% CP overhead). There are $U = 1000$ users each having the arrival rate of 100 packets per second, i.e., $p_j = 0.1$, $j = 1, \cdots, U$. We use the pathloss model $\beta_j = G_0 d_j^{-\alpha}$, where $G_0 = 0.1$, $\alpha = 4$ and d_j is given by

$$
d_j = R_{min} + \frac{(R_{max} - R_{min})\,j}{U} \tag{11}
$$

with $R_{min} = 10$, $R_{max} = 100$. This pathloss model reflects the BS located at the origin and the users are located uniformly along the line $[R_{min}, R_{max}]$. All of users have the same transmit energy constraint, $E_j = E$ for all j. According to the simulation setting, the received signal energy of the worst-case user at the BS is $0dB$ when $E = 70dB$ energy is equally spread over the symbols in a sub-frame. Note that by assuming $-174dBm/Hz$ for the noise spectral density, $E = 90dB$ means only $-3dBm$ ($0.5mW$) per sub-frame in this simulation setting. For comparison, a grant-free multiple access scheme with a random user grouping at a moderately selected value of Q, equal power allocation during the training

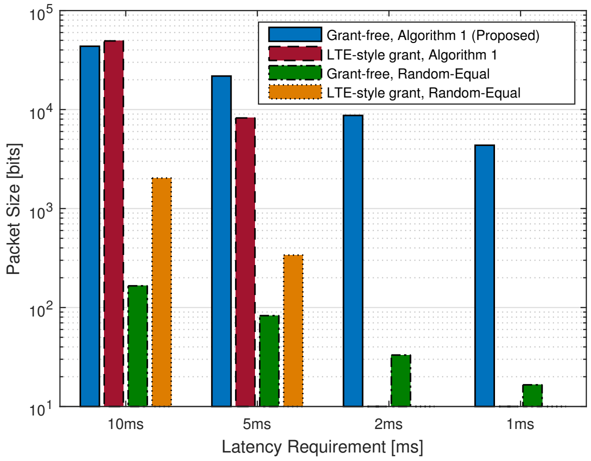

Fig. 3. The allowed packet size and the latency requirement when $M = 64$, $E = 90dB$, and the corresponding outage probability is 10^{-3}.

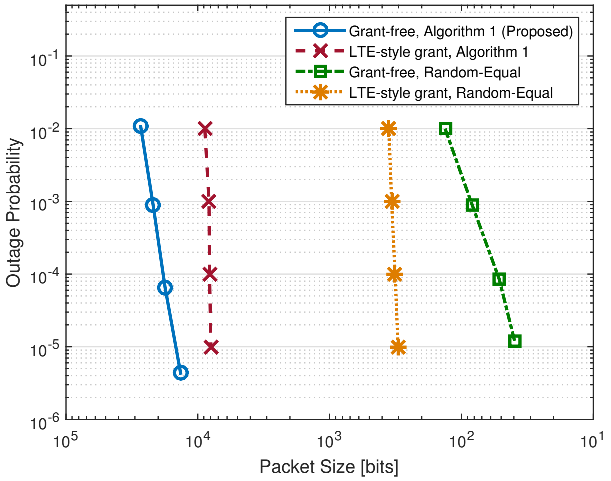

Fig. 4. The allowed packet size and the corresponding outage probability when $M = 64$, $E = 90dB$, and the latency requirement is $5ms$.

and data transmission phases is assumed as the conventional scheme. Two LTE-style grant-based multiple access systems, one with the latency-optimal scheduling in [11], and the other with a random scheduling are also considered for comparison.

In Fig. 3, the allowed packet size and the latency requirement of the proposed scheme is compared to that of conventional schemes as the latency requirement changes from $10ms$ to $1ms$ when $M = 64$, $E = 90dB$, and the reliability requirement is 10^{-3}. Here, we assume that the delay caused by the LTE-style grant procedure requries 4 times of the subframe length ($4ms$) so that only the remaining portion of the latency requirement can be used for data transmission. From the results, it is shown that, i) the proposed grant-free multiple access scheme outperforms the LTE-style grant-based scheme with a random scheduling, ii) althogh the LTE-style grant-based scheme can be improved by adopting the optimal scheduling policy, the proposed grant-free multiple access scheme is better in a low-latency regime, and iii) the proposed delay-optimal resource allocation is critical in a grant-free multiple access scheme.

In Fig. 4, the allowed packet size and the corresponding average outage probability of the proposed scheme is compared to that of conventional schemes as the reliability requirement changes from 10^{-2} to 10^{-5} when $M = 64$, $E = 90dB$, and the latency requirement is $5ms$. From the results, it is shown again that, although the conventional system suffers from the inefficiency caused by the latency and reliability requirements, the proposed system can improve the spectral efficiency significantly.

V. Concluding Remark

In this paper, a grant-free multiple access scheme is proposed for URLLC with a sporadic traffic characteristic in a large-scale antenna systems. The proposed scheme maximizes the allowed packet size while maintaining given latency and reliability requirements so that the spectral efficiency of a URLLC can be greatly improved. Thus, if combined with a good admission control scheme, it can be a promising candidate of the multiple access scheme for URLLC.

VI. Acknowledgement

This work was supported by ICT R&D program of MSIP/IITP. {B0126-16-1012, Multiple Access Technique with Ultra-Low Latency and High Efficiency for Tactile Internet Services in IoT Environments}

References

[1] R. N. Mitra and D. P. Agrawal, "5G mobile technology: A survey," *ICT Express*, Elsevier, 2015, vol. 1, no. 3, pp. 132–137.

[2] ITU-R, "IMT vision - framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation ITU-R M.2083-0, September 2015.

[3] NGMN, "5G use cases, deployment scenarios and framework of requirements," *RAN 5G Workshop - the Start of Something*, Phoenix, AZ, U.S.A., September 19, 2015.

[4] J.H. Kim et al, "Radio Access Technology Requirements For the Ultra-Reliability and Low Latency Services," *KICS 2016 Winter*, Gangwon-Province, South Korea, January 21, 2016.

[5] N. Brahmi, O.N.C. Yilmaz, K.W. Helmersson, S.A. Ashraf, and J. Torsner, "Deployment strategies for ultra-reliable and low-latency communication in factory automation," *IEEE Global Commun. Confer. (Globecom) URLLC workshop*, San Diego, CA, U.S.A., December 2015.

[6] LTE-A book

[7] N.A. Johansson, Y.-P.E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," *IEEE Inter. Confer. Commun. (ICC) workshop on 5G & beyond - enabling technologies and applications*, London, U.K., June 2015.

[8] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," *Inter. Symp. Wireless Commiun. Syst. (ISWCS)*, Barcelona, Spain, August 2014.

[9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 35903600, November 2010.

[10] P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, and A. Ugolini, "Modulation formats and waveforms for 5G Networks: who will be the heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 80–93, Nov 2014.

[11] K.J. Choi and K.S. Kim "Latency-Optimal Uplink Scheduling Policy in Training-based Large-Scale Antenna Systems," *submitted to IEEE Trans. Wireless Commun.* available: http://arxiv.org/abs/1607.07547

[12] A. Schrijver, *Theory of Linear and Integer Programming*. New York, Wiley, 1986.