IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2019, accepted November 14, 2019, date of current version November 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954203

Performance Analysis of Grant-Free Multiple Access for Supporting Sporadic Traffic in Massive IoT Networks

TAEHOON KIM[1], (Member, IEEE), AND BANG CHUL JUNG[2], (Senior Member, IEEE)
[1]Agency for Defense Development, Daejeon 34186, South Korea
[2]Department of Electronics Engineering, Chungnam National University, Daejeon 34134, South Korea

Corresponding author: Bang Chul Jung (bcjung@cnu.ac.kr)

This work was supported in part by the NRF through the Basic Science Research Program funded by the Ministry of Science and ICT (MSIT) under Grant NRF2019R1A2B5B01070697, in part by the MSIT, South Korea, under the Information Technology Research Center (ITRC) support program under Grant IITP-2019-2017-0-01635 supervised by the Institute for Information and Communications Technology Promotion (IITP), and in part by the Institute of Information and Communications Technology Planning and Evaluation grant funded by the Korea Government (MSIP, Research on Near-Zero Latency Network for 5G Immersive Service) under Grant 2015-0-00278.

ABSTRACT Grant-free multiple access (GFMA) protocol has been regarded as a key element to support sporadic traffic generated from massive internet-of-things (IoT) networks. In GFMA protocol, each IoT device transmits data packets without grant from a base station (BS) via pre-reserved uplink resources. Packet collisions inherently occur when multiple IoT devices transmit packets by using the same radio resource, but the collision effect can be alleviated with multi-packet reception (MPR) capability of the BS. Since a number of studies have focused on improving the physical layer performance such as bit error rate, they may be hard to provide intuitions from the MAC layer perspective when a number of IoT devices sporadically generate uplink packets and attempt the GFMA. In this paper, we thoroughly investigate the GFMA from the MAC layer perspective. We provide an analytical framework based on a Markov chain to capture the performance of the GFMA in terms of packet transmission success probability, ergodic throughput, and access delay. Through simulations, we validate our analytical framework and verify the necessity of adopting MPR technique for supporting a massive number of IoT devices generating sporadic traffic.

INDEX TERMS Cellular IoT networks, grant-free multiple access, sporadic traffic, massive connectivity, multi-packet reception.

I. INTRODUCTION

Internet-of-things (IoT), which connects a massive number of IoT devices with a wide range of applications through IP-based networks, has been considered as a key enabler for Industry 4.0 [1]. Due to the advantage of cellular networks such as coverage and security, the cellular networks have attracted great attention as one of candidates for implementing IoT. Accordingly, there have been a number of studies for implementing IoT in practical cellular networks such as LTE/5G new radio (NR) [2], [3].

A number of IoT devices are expected to sporadically transmit small-sized packets in uplink direction for reporting purpose [4]. In this case, each IoT device may transit to a sleep mode after completion of packet transmissions for saving energy consumption and release its connection with the base station (BS) [5]. This implies that each IoT device should perform a random access (RA) procedure to (re-)establish the connection with the BS when it has a new packet to be sent to the IoT server.

The RA procedure consists of 4-steps of handshaking [6], which takes several tens of millisecond (ms) [7]. From the perspective of data packet transmission, the RA procedure can be regarded as additional signaling procedure required in advance. In particular, it has been considered as a critical signaling overhead for supporting sporadic traffic generated from IoT devices. As the size of packet becomes smaller, this signaling procedure becomes more inefficient. Without considering significant modifications of the legacy protocol, minimizing the RA delay spent before the actual data transmission as small as possible can be the straightforward approach to improve the latency performance. For the last

The associate editor coordinating the review of this manuscript and approving it for publication was Zubair Fadlullah.

few years, a number of studies have focused on reducing the RA delay by mitigating collision problem [7]–[9], however, the latency cannot be reduced below a certain value due to the inherent handshaking procedure.

One of the solutions introduced in [10] is contention-based uplink transmissions. Andreev *et al.* [11] proposed a contention-based access, where each IoT device sends data packets over the shared resources without spending time on the RA procedure, i.e., without grant acquisition procedure. This may reduce the latency, however, a collision problem may still occur when multiple IoT devices utilize the same resource.

To address the collision problem, non-orthogonal transmission mechanisms [12], such as a sparse code multiple access (SCMA) [13], a multi user shared access (MUSA) [14], a pattern division multiple access (PDMA) [15], and a resource spread multiple access (RSMA) [16], are expected to be applied on top of the contention-based framework [17]. The MPR capability provided by such non-orthogonal transmission mechanisms can mitigate the occurrence of packet collisions. In line with those of approaches, enormous efforts have been made to improve the grant-free protocol. Wang *et al.* [18] proposed a compressive sensing based access mechanism using the sparsity of sporadic traffic. Similarly, Zhang *et al.* [19] studied the joint user activity and data detection problem in an uplink grant-free non-orthogonal multiple access (NOMA) system based on compressive sensing. Dogan *et al.* [20] proposed NOMA with index modulation for grant-free access. Moreover, Berardinelli *et al.* [21] analyzed the reliability of the contention-based access protocol.

To the authors' best knowledge, most of previous studies have still focused on the physical layer performance, but did not provide any insightful intuitions from the MAC layer perspective. In this paper, we thoroughly investigate grant-free multiple access (GFMA) from the MAC layer perspective under the IoT scenario with sporadic traffic. The main contributions of this paper can be summarized as follows:

- We propose an analytical framework based on a Markov chain to capture the MAC layer performance of the GFMA such as packet transmission success probability, ergodic throughput, and access delay, and validate our analytical framework through extensive system-level simulations.
- We investigate the effect of resource configuration on the MAC layer performance, and verify that resource-spreading approach is much beneficial than multi-channel approach to support the IoT scenario with sporadic traffic.
- We investigate the effect of sporadic traffic from a large number of IoT devices on the system, and find that even though the IoT scenario with sporadic traffic imposes critical burden on the system, a few more resources are sufficient enough to support it.

The rest of this paper is organized as follows. In Section II, we describe the GFMA protocol in more detail. In Section III, we analyze the GFMA with a Markov chain model from the MAC layer perspective. In Section IV, we provide numerical results. Finally, we draw conclusions in Section V.

II. GRANT-FREE MULTIPLE ACCESS

In this section, we explain the system model and introduce the GFMA protocol in more detail.

A. SYSTEM MODEL

In this subsection, we first explain the system model we considered in this paper. Fig. 1 describes the system model. We consider a single cell, and the BS is assumed to adopt the MPR-capable receiver. We consider a general IoT scenario [7]–[9], where each IoT device sporadically generates its data packet. Accordingly, we use a Poisson process as a traffic model [4].[1] In other words, each IoT device generates a new packet following a Poisson distribution with arrival rate of λ [22], and its probability mass function (pmf) can be expressed as

$$\Pr\{K = k\} = p_k = \frac{e^{-\lambda T_{\mathrm{P}}} (\lambda T_{\mathrm{P}})^k}{k!}, \tag{1}$$

where K represents the number of newly arrived packets within a given time interval T_{P}. It is worth noting that the packet arrival rate λ is assumed to have an arbitrary small value due to the sporadic characteristic of the traffic generated from the IoT devices [8]. We assume that the generated packet is sufficiently small and thus it can be successfully transmitted through each transmission attempt if there is no packet collision. If there exists at least a packet to be sent in each IoT device's queue (or, equivalently, buffer), it attempts uplink transmissions with a probability of p_t on the next available GFMA resource, which is periodically reserved for the GFMA protocol only. Note that the transmit probability p_t plays an important role to control the overall traffic load in the system. Here, the period of the GFMA resource is denoted as T_{p}. Let N_{R} denote the amount of pre-reserved GFMA resources per period, which can be defined as $N_{\mathrm{R}} = N_{\mathrm{C}} \times N_{\mathrm{M}}$, where N_{C} and N_{M} represent the number of channels and the MPR capability of the BS, respectively. N_{C} and N_{M} can be arbitrary values.

B. OVERALL PROCEDURE

In this subsection, we introduce the overall procedure of the GFMA protocol, which consists of 2-steps of handshaking procedure. Key idea of the GFMA is that whenever each IoT device has packets, it attempts uplink transmissions with a probability of p_t at the next available GFMA resource. Fig. 2 visualizes the handshaking procedure of the GFMA protocol, where delay-incurring components are also specified. It is worth noting that the time duration of each component takes

[1] The IoT scenario can be categorized into twofolds [4]: One is the overload scenario due to the event-driven situation such as a simultaneous power-on event after a disaster situation, which can be modeled by either beta distribution or uniform distribution. The other is the general scenario without traffic overload, which can be modeled by Poisson distribution.

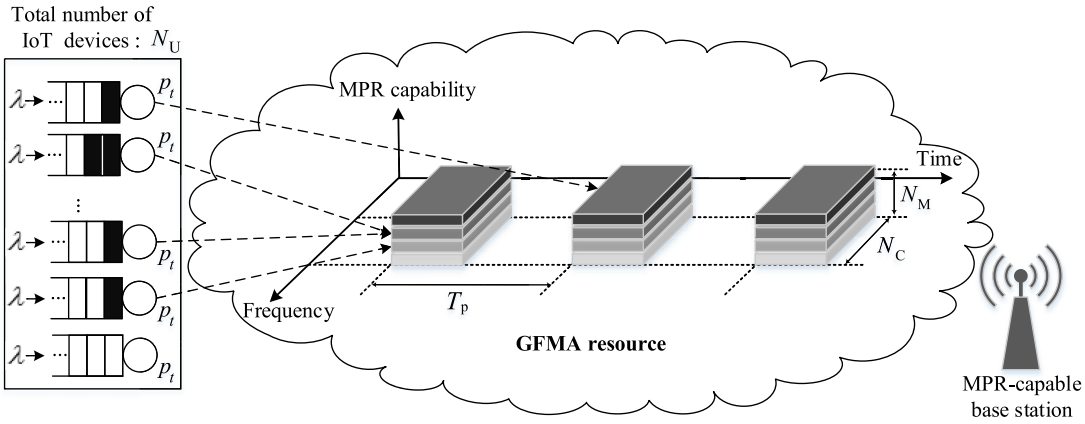


FIGURE 1. System model where active IoT devices among N_U IoT devices attempt its uplink packet transmissions with a probability of p_t at the next-available GFMA resource. T_P represents the period of the GFMA resource, and N_C and N_M represent the number of channels and the MPR capability of the BS, respectively, which satisfy $N_R = N_C \times N_M$.

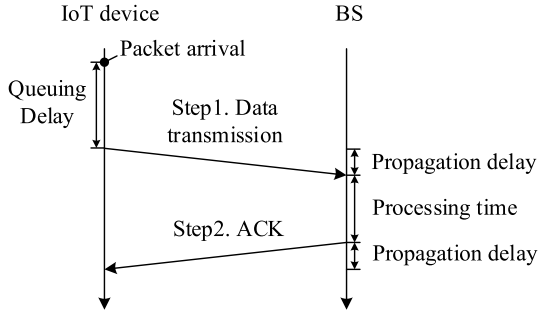


FIGURE 2. Overall procedure of the GFMA protocol.

an integer multiple of the transmission time interval (TTI) value (e.g., 1ms in LTE/LTE-A systems [23]). The details of each step are as follows:

- **(Step 1) Packet transmissions** : Each IoT device randomly selects a channel among N_C channels within the GFMA resource, and transmits its data packet with a probability of p_t. For contention resolution, whenever each IoT device transmits the packet, it starts a contention resolution (CR) timer.
- **(Step 2) Acknowledgement** : The BS attempts to decode the packets received through the GMFA resource. The BS transmits the acknowledgement (ACK) messages to the IoT devices, whose transmitted packets are successfully decoded. If each IoT device receives the ACK message before the CR timer expires, then it regards the packet transmission as a success. Otherwise, it regards the packet transmission as a failure and repeats Step 1 and Step 2 for reattempting uplink packet transmission.

C. DISCUSSIONS

From Fig. 2, we can conjecture the overall delay when the packet is successfully transmitted at once without any packet collision, which can be calculated as the summation of the queuing delay, the propagation delays, and the processing time. When the packet experiences the collision, both Step 1 and Step 2 should be repeated until the packet is successfully transmitted. In this case, the time duration spent during the reattempts should be considered for the overall delay. The detailed explanation on the success of packet transmission and the overall delay will be followed in the next section.

III. PERFORMANCE ANALYSIS

In this section, we propose a Markov chain to capture the performance of the GFMA protocol in terms of packet transmission success probability, ergodic throughput, and access delay. Moreover, we mathematically analyze those of performance metrics in detail.

A. MARKOV CHAIN MODEL

Fig. 3 shows the state transition diagram of the proposed Markov chain. Our approach is proposed based on a well known literature [24] and we tailor it for our system model. Thus, each state represents the queue length of each IoT device, where the queue length implies the number of packets in the queue. It is noteworthy that since our proposed Markov chain model considers the queue length of each IoT device as state, it has primary benefit that it can be applied to any traffic model not restricted to Poisson based traffic model. The state space $\mathcal{S}$ can be defined as

$$\mathcal{S} = \{0, 1, \ldots, i, \ldots, I\}, \tag{2}$$

where I represents the maximum queue size. $\pi_i(n)$ represents the state probability that the queue length equals to i at time n and thus the distribution of the state probability at time n, $\boldsymbol{\pi}(n)$, can be denoted as

$$\boldsymbol{\pi}(n) = \{\pi_0(n), \pi_1(n), \ldots, \pi_I(n)\}. \tag{3}$$

A state transition occurs whenever each IoT device attempts to perform an uplink transmission, and the state

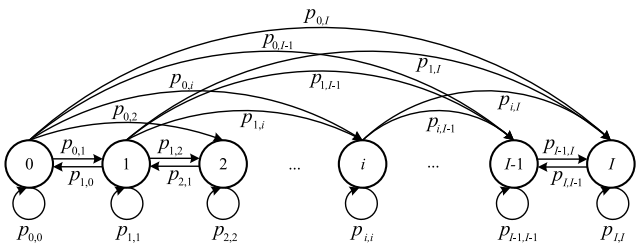


FIGURE 3. A state transition diagram of the proposed Markov chain, where $p_{j,k}$ represents the state transition probability from state j to state k.

transition probability from state j to state k, $p_{j,k}$, is given in (4), shown at the next page, where p_s represents the transmission success probability. Even though each IoT device does not perform actual transmissions due to p_t value, the state update occurs. Note that our proposed Markov chain model can also describe a drastic packet generation (e.g., $p_{j,k}$ where $k \geq j$), but this hardly occurs in the general IoT scenario with sporadic traffic.

$$p_{j,k} = \begin{cases} p_k, j = 0, \\ \quad 0 \leq k \leq I \\ p_{k-j}p_t(1-p_s) + p_{k-j}(1-p_t) + p_{k-j+1}p_t p_s, \\ \quad 1 \leq j \leq I-1, \quad j < k \leq I \\ p_0 p_t p_s, \\ \quad 1 \leq j \leq I, \quad k = j-1 \\ p_0 p_t(1-p_s) + p_0(1-p_t) + p_1 p_t p_s, \\ \quad 1 \leq j \leq I-1, \quad k = j \\ 1 - p_0 p_t p_s, \\ \quad j = I, \quad k = j \end{cases} \tag{4}$$

B. PACKET TRANSMISSION SUCCESS PROBABILITY

Packet transmission success probability, p_s, is defined as the probability that a packet is successfully transmitted when each IoT device performs an uplink transmission. In order to mathematically derive the packet transmission success probability, we should obtain the probability of existing at least a packet in the queue, τ, which implies the probability that attempts to perform the uplink transmissions. In order to derive τ, we should obtain a stationary distribution, $\boldsymbol{\pi}$.

The steady-state probabilities can be expressed as

$$\pi_0 = \pi_0 p_{0,0} + \pi_1 p_{1,0}, \tag{5}$$

$$\pi_i = \sum_{j=0}^{i+1} \pi_j p_{j,i}, \quad i \in [1, I-1], \tag{6}$$

and

$$\pi_I = \sum_{j=0}^{I-1} \pi_j p_{j,I} + \pi_I \left(1 - p_{I-1,I}\right). \tag{7}$$

With above equations, it is hard to express the stationary distribution in a closed-form. Hence, we obtain the stationary distribution by using the fact that it converges to a certain distribution regardless of the initial state probabilities.

Algorithm 1 Finding p_s and τ

Input : N_U, λ, p_t, T_P, N_C, N_M, I, $\gamma_{\max}$ and ϵ
Output : p_s, τ

```
γ ← 0
while (1) do
  if extremely sporadic traffic then
    Calculate π_0
  else
    Generate a one-step transition matrix, P
    Find a stationary distribution, π = {π_0, ..., π_I}
  end if
  Calculate τ = 1 − π_0 // from (9)
  Calculate p_s // from (10)
  if abs (p_s^prev − p_s) < ε then
    γ ← γ + 1
    if γ == γ_max then
      break
    else
      γ ← 0
    end if
  end if
  p_s^prev ← p_s
end while
```

Let $\mathbf{P}$ denote the one-step state transition matrix, which has $p_{j,k}$ as an element of the j-th row and the k-th column. The steady-state probability of the state i, π_i, can be obtained by $\pi_i = \lim_{n\to\infty} \pi_i(n)$. Finally, the stationary distribution, $\boldsymbol{\pi}$, which is denoted as $\boldsymbol{\pi} = \{\pi_i\}$ for $i \in [0, I]$, can be derived by:

$$\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{\pi}(n) = \lim_{n\to\infty} \mathbf{P}^n \boldsymbol{\pi}(0), \tag{8}$$

where $\boldsymbol{\pi}(0)$ represents the distribution of the initial state probabilities.

Now, we have

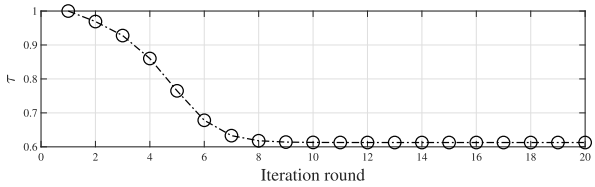
$$\tau = 1 - \pi_0, \tag{9}$$

which is a function of p_s, since $\boldsymbol{\pi}$ and $\boldsymbol{P}$ are also functions of p_s (see (4) and (8)). From the viewpoint of system, p_s can be derived as

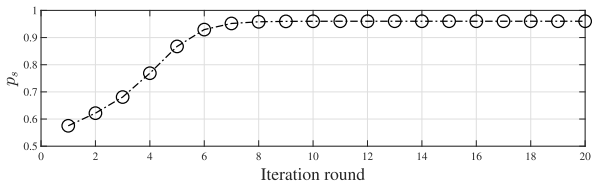
$$p_s = \sum_{k=0}^{N_M-1} \binom{N_U - 1}{k} \left(\frac{\tau p_t}{N_C}\right)^k \left(1 - \frac{\tau p_t}{N_C}\right)^{N_U-1-k}, \tag{10}$$

which is a function of τ. Finally, (9) and (10) comprise a non-linear equation and thus τ and p_s can be found by numerical methods. Our numerical approach is summarized as Algorithm 1.

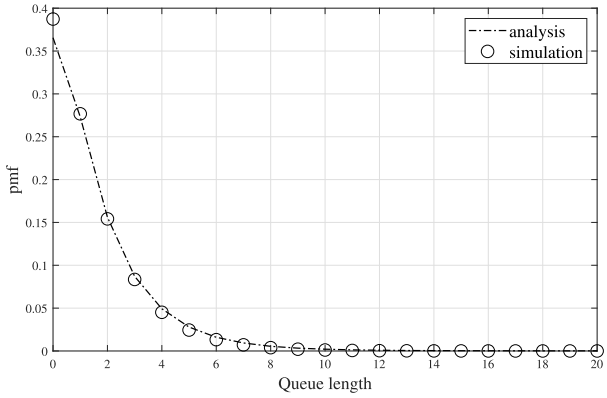
Fig. 4 shows the convergence of the solutions based on Algorithm 1. As the iteration round proceeds, both τ and p_s interact with each other and consequently converge to the final values within a few iteration rounds. Since we set the $\gamma_{\max}$ value as 20 in this example, we can find that the algorithm is terminated when 20 samples of p_s^{prev} - p_s maintain



(a) Convergence of τ according to the iteration round



(b) Convergence of p_s according to the iteration round



(c) Stationary distribution of the Markov chain

FIGURE 4. Finding τ, p_s, and π when $N_U = 15$, $\lambda = 0.3$, $p_t = 0.52$, T_P=10ms, $N_C = 1$, $N_M = 8$, $I = 20$, $\gamma_{max} = 20$, and, $\epsilon = 0.001$.

arbitrary small value (see Fig. 4(a) and Fig. 4(b)). Finally, Fig. 4(c) shows the stationary distribution of the Markov chain, in which the IoT device can maintain the queue length less than 4 with a probability of 0.9.

Especially for extremely sporadic traffic, it is reasonable to assume that $p_k = 0$, for $k \geq 2$, and, thus, p_1 can be expressed as $1-p_0$. In this case, we can express the stationary distribution in a closed-form. The steady-state probabilities can be expressed as

$$\pi_0 = \pi_0 p_{0,0} + \pi_1 p_{1,0}, \tag{11}$$

$$\pi_i = \pi_{i-1} p_{i-1,i} + \pi_i p_{i,i} + \pi_{i+1} p_{i+1,i}, \quad i \in [1, I-1], \tag{12}$$

and

$$\pi_I = \pi_{I-1} p_{I-1,I} + \pi_I \left(1 - p_{I,I-1}\right) \tag{13}$$

With above equations, we have

$$\pi_i = \begin{cases} \pi_0, & i = 0 \\ \pi_0 \prod_{j=0}^{i-1} \left(\frac{p_{j,j+1}}{p_{j+1,j}}\right), & i \in [1, I]. \end{cases} \tag{14}$$

By using the normalization condition for the stationary distribution, we have

$$1 = \sum_{i=0}^{I} \pi_i = \pi_0 \left(1 + \sum_{i=1}^{I} \prod_{j=0}^{i-1} \left(\frac{p_{j,j+1}}{p_{j+1,j}}\right)\right). \tag{15}$$

Therefore, by using (15), π_0 can be expressed as

$$\pi_0 = \left(1 + \sum_{i=1}^{I} \prod_{j=0}^{i-1} \left(\frac{p_{j,j+1}}{p_{j+1,j}}\right)\right)^{-1}. \tag{16}$$

Similarly, after substituting (16) in (9), we can also find τ and p_s by using numerical methods (see also Algorithm 1).

C. ERGODIC THROUGHPUT

Ergodic throughput, T, is defined as the total number of IoT devices which successfully transmit their packet during a single period of the GFMA resource, which can be derived as

$$T = \min\Big(\underbrace{N_U \times \lambda}_{\text{offered load}} , \underbrace{N_U \times \overbrace{p_t p_s}^{\text{function of } p_t}}_{\text{system capacity}} \Big), \tag{17}$$

where the system capacity implies the maximum throughput that the system can achieve using the given amount of the GFMA resources. Note that even though the amount of the GFMA resources is given, how to set p_t can vary the system capacity. Therefore, for a given p_t, the system may operate in a saturated condition or a unsaturated condition. When offered load exceeds the system capacity, the system operates in a saturated condition, which implies that all IoT devices have at least a packet to transmit. On the contrary, offered load is less than the system capacity, the system operates in a unsaturated condition, which implies that some IoT devices have at least a packet to transmit. Using (17), we can derive the maximum packet arrival rate that the system can support, λ_{max}, as follows:

$$\lambda_{max} = \max_{p_t} p_t p_s . \tag{18}$$

D. ACCESS DELAY

Access delay, D, is defined as the time duration from a new packet arrival to the successful completion of transmitting the corresponding packet, which can be expressed as

$$D = D_Q + D_T, \tag{19}$$

where D_Q and D_T represent the queuing delay and the transmission delay, respectively.

Queuing delay is defined as the time duration between the new packet arrival and the first attempt of uplink transmission, which can be derived as

$$D_Q = \sum_{i=0}^{I} \pi_i D_{Q,i}, \tag{20}$$

where $D_{Q,i}$ represents the queuing delay of a new packet which is arrived to the queue with the length of i, which can

be expressed as

$$D_{Q,i} = \frac{T_p}{2} + i \cdot T_p \cdot \mathbb{E}_X[X], \quad i \in [0, I] \tag{21}$$

where X represents the number of transmission opportunities until the packet is successfully transmitted, which follows a geometric distribution with parameter $p_t p_s$ [22], and its pmf can be expressed as

$$f_X(x) = (1 - p_t p_s)^{x-1} (p_t p_s) \quad \text{for } x \in \{1, 2, 3, ...\}. \tag{22}$$

Transmission delay is defined as the time duration from the first attempt of uplink transmission to the time when the corresponding ACK message is successfully received, which can be expressed as

$$D_{\mathrm{T}} = \mathbb{E}_X[(X-1)T_{\mathrm{p}} + T_{\mathrm{S1-S2}}], \tag{23}$$

where $T_{\mathrm{S1-S2}}$ represents the time duration from Step 1 to Step 2, which can be expressed as $T_{\mathrm{S1-S2}} = 2 \times T_{\mathrm{prop}} + T_{\mathrm{proc}}$ as shown in Fig. 2. Here, T_{prop} and T_{proc} represent the propagation delay via air-interface and the processing time at the BS, respectively.

It is noteworthy that when the system operates in a saturated condition, the queuing delay may infinitely increase. Even though the system can minimize the transmission delay by adjusting p_t, however, the access delay may also infinitely increase due to the large impact of the queuing delay. As a result, in order to achieve meaningful latency performance during the packet transmissions, the system should be operated in a unsaturated condition. Furthermore, in order to support low-latency featured data transmissions (e.g., ∼9ms in our system model), p_t should be 1. This will be further discussed in Table 2. The system should utilize sufficient enough GFMA resources, and carefully select the optimal transmit probability, $p_t{}^*$, which can minimize the access delay. We can formulate an optimization problem as

$$\begin{aligned} p_t{}^* \;=\; & \operatorname{argmin} D(p_t) \\ & \text{subject to } p_t \in [0, 1]. \end{aligned} \tag{24}$$

Using any optimization algorithms, we can obtain $p_t{}^*$.

IV. NUMERICAL RESULTS

We perform system-level simulations using a process-oriented discrete-event simulation package, CSIM, and specific simulation parameters are listed in Table 1. We assume that there is no frame error due to the channel condition, and also assume that the BS successfully decodes packets whenever less than N_{M} devices utilize the same resource, where N_{M} represents the MPR capability of the BS. In all figures, markers and lines indicate the simulation and analytical results, respectively.

From Fig. 5 to Fig. 9, we first investigate the effect of resource configuration on the performance. Moreover, in Table 2, we then investigate the effect of sporadic traffic generated from a massive number of IoT devices on the performance. The delay performance is evaluated under the assumption that the transmission time interval (TTI) value is 1ms and thus the point what we should note is that the minimally achievable access delay is 9ms.[2]

TABLE 1. Simulation parameters and values [6], [25].

Parameters	Values
Physical layer related parameters and values	
Transmission time interval (TTI)	1ms
Propagation delay (T_{prop})	1ms
Processing time at the BS (T_{proc})	1ms
Time duration for handshaking ($T_{\mathrm{S1-S2}}$)	4ms
MAC layer related parameters and values	
Period of GFMA resource (T_{P})	10ms
Amount of radio resources per T_{P} (N_{R})	8, 24
Number of IoT devices (N_{U})	15,000, 1,500, 150, 15
Packet arrival rate per T_{P} (λ)	0.0003, 0.003, 0.03, 0.3
Transmit probability (p_t)	$0 \sim 1$
Contention resolution (CR) timer	8ms
Algorithm related parameters and values	
Queue length (I)	20
Required number of comparisons ($\gamma_{\max}$)	20
Termination condition (ϵ)	0.001

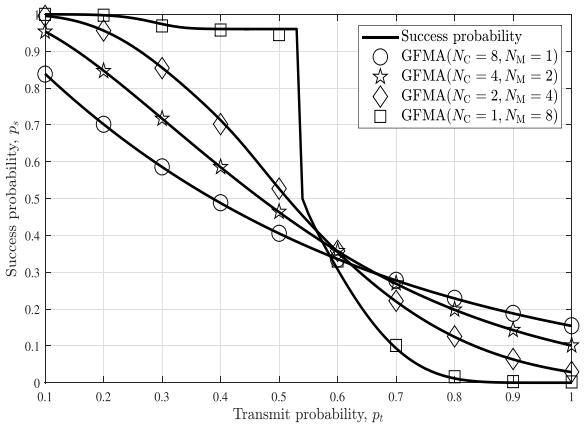


FIGURE 5. Comparison of success probability for varying values of p_t in various combinations of N_{C} and N_{M} when $N_{\mathrm{U}} = 15$ and $\lambda = 0.3$.

Fig. 5 shows the transmission success probability for varying the transmit probability, p_t. The success probability tends to decrease as p_t increases for all combinations of N_{C} and N_{M} due to packet collisions. Especially when the MPR capability is sufficient enough, the higher value of success probability can be achieved when p_t is carefully adjusted, e.g., $p_t < 0.53$ in case that $N_{\mathrm{C}} = 1$ and $N_{\mathrm{M}} = 8$, compared to other combinations. The resource configuration of $N_{\mathrm{C}} = 1$ and $N_{\mathrm{M}} = 8$ implies that it cannot support more than 8 devices at the same time. This results in a steep decrease of p_s as the traffic load increases, which can be observed when $p_t > 0.53$.

Fig. 6 shows the ergodic throughput for varying the transmit probability, p_t. The ergodic throughput decreases as p_t decreases, since each IoT device hardly transmits its packet

[2]Queuing delay (D_{Q}) and the transmission delay (D_{T}) are 5ms and 4ms, respectively. In near future, we expect that the advances in hardware and air-interface can further contribute to improve the delay performance.

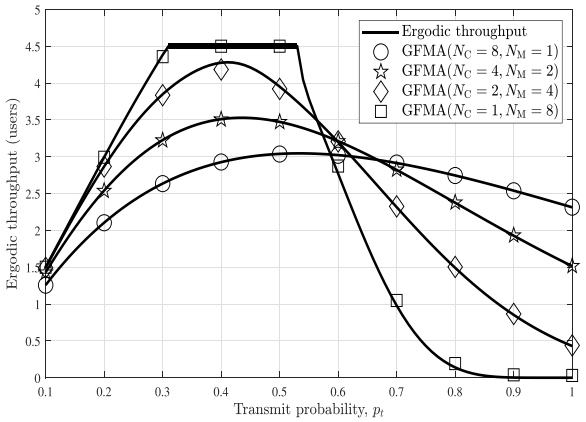


FIGURE 6. Comparison of ergodic throughput for varying values of p_t in various combinations of N_C and N_M when $N_U = 15$ and $\lambda = 0.3$.

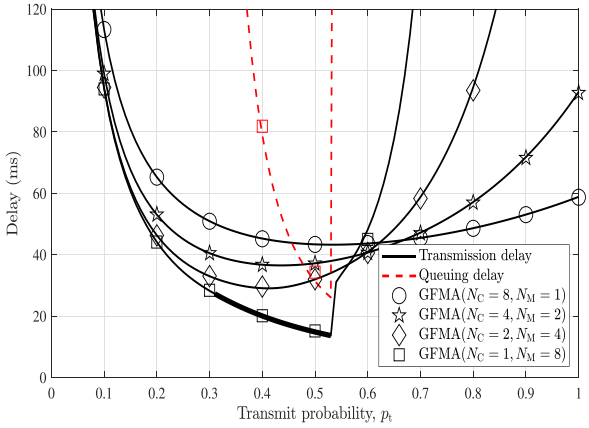


FIGURE 7. Comparison of transmission and queuing delay for varying values of p_t when $N_U = 15$ and $\lambda = 0.3$.

due to low p_t. It also decreases as p_t increases, since each IoT device hardly succeeds in its transmissions due to low p_s (or, equivalently, due to high packet collision probability). Accordingly, there exists an optimal p_t^*, which varies according to the combinations of N_C and N_M. With the optimal p_t^*, the case that $N_C = 1$ and $N_M = 8$ achieves the highest throughput. It is noteworthy that the throughput of the case that $N_C = 1$ and $N_M = 8$ becomes unchanged when $0.31 < p_t < 0.53$ since all input traffic is transferred to the BS in this range, which is also called unsaturated case. p_t should be carefully adjusted for the optimal throughput performance.

Fig. 7 shows both the transmission delay and the queuing delay for varying values of p_t. The transmission delay increases as p_t decreases, since each IoT device hardly attempts its uplink transmissions due to low p_t and thus each IoT device spends time to reattempt uplink transmissions. As p_t increases, the transmission delay also increases, since each IoT device hardly succeeds in its transmissions due to low p_s (see Fig. 5). Thus, each combination has an optimal p_t^* that minimizes the transmission delay as in ergodic throughput case. With p_t^*, the case that $N_C = 1$ and $N_M = 8$ achieves the best transmission delay. Basically, queuing delay

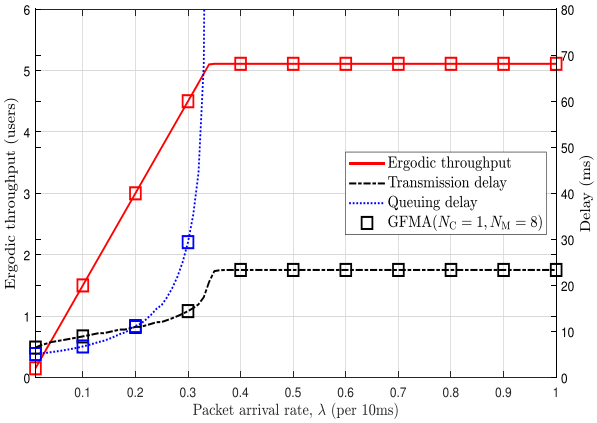


FIGURE 8. Ergodic throughput, queuing delay, and transmission delay of the GFMA with p_t^* for varying values of λ when $N_U = 15$, $N_C = 1$, and $N_M = 8$.

is infinite if service rate is lower than the arrival rate. In our case, the service rate depends on the ergodic throughput. In Fig. 7, only the case that $N_C = 1$ and $N_M = 8$ operates in a unsaturated condition in which the queuing delay is finite. As shown in Fig. 7, the queuing delay drastically changes according to p_t and thus p_t should be carefully chosen by considering both the transmission and queuing delay. The case that $N_C = 1$ and $N_M = 8$ can provide meaningful latency performance by adjusting p_t value, but it cannot provide low-latency performance, i.e., ~9ms. To achieve the lowest latency performance, the system should be operated with $p_t = 1$, which implies the more radio resources are required. With more radio resources, the delay monotonically decreases as p_t increases. This issue will be discussed in detail in Table 2.

Fig. 8 shows the relationship among the ergodic throughput, queuing delay, and transmission delay of the GFMA with optimal transmission probability in (24) for varying the packet arrival rate, λ, when $N_U = 15$, $N_C = 1$, and $N_M = 8$. The ergodic throughput becomes saturated beyond a certain point, i.e., $\lambda = 0.32$, which is the maximum load that the system can support. In other words, the system operates in a unsaturated condition whenever $\lambda \leq 0.32$, and it operates in a saturated condition whenever $\lambda > 0.32$. Interestingly, the transmission delay has also similar trend with the ergodic throughput. On the contrary, the queuing delay exponentially increases until the system is saturated. It is worth noting that the access delay increases to infinity because of the queuing delay when the system is saturated.

Fig. 9 shows the access delay for varying values of λ when $N_U = 15$. For each λ, optimal transmit probability p_t^* is applied. The access delay gradually increases for low traffic load, but it drastically increases as the traffic load approaches to the saturated condition. When the required access delay is set to 20ms, the supportable traffic load in the case that $N_R = 16$ is equal to 0.7, while the supportable traffic load the case that $N_R = 8$ is equal to 0.2. It is noteworthy that nearly 3.5 times more traffic can be supported by increasing resources 2 times, which comes from the statistical multiplex effect.

TABLE 2. Summary of performance of the GFMA in various scenarios when $T_P = 10$ms and $T_{S1-S2} = 4$ms.

Scenario	N_U	λ	N_R	N_C	N_M	p_t^*	Success probability (%)	Ergodic throughput	Queueing delay (ms)	Transmission delay (ms)	Access delay (ms)
Scenario 1	150	0.03	8	8	1	0.05	39.29	2.95	∞	502	∞
			8	4	2	0.043	52.13	3.38	∞	437	∞
			8	2	4	0.040	66.30	3.91	∞	373	∞
			8	1	8	0.043	80.02	4.50*	2,598	284	2,882
	1,500	0.003	8	8	1	0.0053	37.03	2.94	∞	5088	∞
			8	4	2	0.0043	52.16	3.36	∞	4,452	∞
			8	2	4	0.0040	65.49	3.89	∞	3,811	∞
			8	1	8	0.0042	76.39	4.50*	44,880	3,111	47,991
Scenario 2	15	0.3	24	8	3	1.0	97.92	4.50*	5.21	4.22	9.43
				3	8	1.0	99.99	4.50*	5.20	4.00	9.20
	150	0.03	24	8	3	1.0	97.84	4.50*	5.17	4.22	9.39
				3	8	1.0	99.98	4.50*	5.16	4.00	9.16
	1,500	0.003	24	8	3	1.0	97.76	4.50*	5.02	4.23	9.25
				3	8	1.0	99.98	4.50*	5.02	4.00	9.02
	15,000	0.0003	24	8	3	1.0	97.75	4.50*	5.00	4.23	9.23
				3	8	1.0	99.98	4.50*	5.00	4.00	9.00

(*) denotes that the system operates in a unsaturated condition

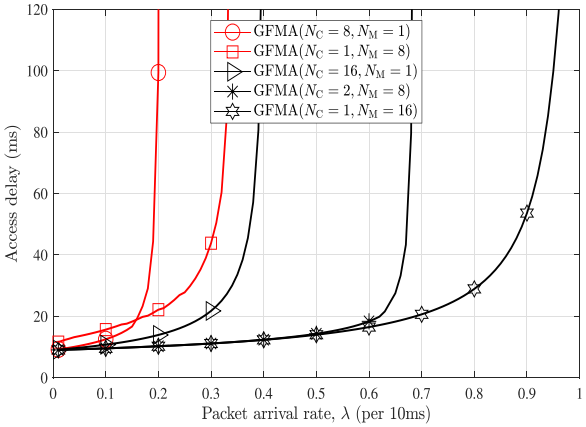


FIGURE 9. Comparison of access delay of the GFMA with p_t^* for varying values of λ in various combinations of N_C and N_M when $N_U = 15$.

From the observation, in order to support much higher load in a unsaturated condition, the system should utilize more resources. Furthermore, if the amount of GFMA resources is the same, increasing N_M is much effective than increasing N_C to support higher load.

Finally, Table 2 summarizes the performance of the GFMA in various scenarios. In Scenario 1, we investigate the effect of N_U on the performance, when the offered load on average is equally set, i.e., $N_U \times \lambda = 4.5$. It is noteworthy that even though the average offered load is the same, sporadically generated packets from a large number of IoT devices give much burden on the system. Scenario 2 shows that a few more resources are sufficient enough to support low-latency transmissions of sporadic traffic generated from a massive number of IoT devices. Furthermore, it is worth noting that the access delay cannot be reduced under a certain value, i.e., 9 ms, even though the system utilizes more resources, since this is an inherent signaling overhead which comes from the handshaking procedure itself as shown in Fig. 2.

V. CONCLUSION

In this paper, we investigated the grant-free multiple access (GFMA) from the MAC layer perspective under the IoT scenario with sporadic traffic. We proposed an analytical framework based on a Markov chain, and mathematically analyzed the MAC layer performance of the GFMA in terms of packet transmission success probability, ergodic throughput, and access delay. Moreover, we thoroughly investigated the effect of MPR capability on the performance, and found the optimal transmit probability which minimizes the delay performance. Through simulations, we validated our analytical framework and verified that the GFMA can support the low-latency transmissions of sporadic traffic generated from a massive number of IoT devices when the MPR capability is sufficient enough.

REFERENCES

[1] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[2] *Service Requirements for Machine-Type Communication*, documnent 3GPP TS 22.368 V13.1.0, Dec. 2014.

[3] *RAN Improvements for Machine-Type Ommunications*, documnent 3GPP TR 37.868 V11.0.0, Oct. 2011.

[4] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in machine-to-machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, Dec. 2016.

[5] T. Kim and I. Bang, "An enhanced random access with preamble-assisted short-packet transmissions for cellular IoT communications," *IEEE Commun. Letter*, vol. 23, no. 6, pp. 1081–1084, Jun. 2019.

[6] *Evolved Universal Terrestrial Radio Access Network(E-UTRAN); Medium Access Control (MAC) Protocol Specification*, documnent 3GPP TS 36.321 V12.7.0, Sep. 2015.

[7] T. Kim, H. S. Jang, and D. K. Sung, "An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1714–1717, Oct. 2015.

[8] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, and J. Y. Ahn, "A novel random access for fixed-location machine-to-machine communications in OFDMA based systems," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1428–1431, Sep. 2012.

[9] E. Park, T. Kim, and Y. Han, "An efficient time-shifted random access scheme for cellular-based IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 522–525, Mar. 2019.
[10] *Study Latency Reduction Techqniues for LTE (Release 13)*, document 3GPP TR 36.881 V0.5.0, Nov. 2015.
[11] S. Andreev, A. Larmo, M. Gerasimenko, V. Petrov, O. Galinina, T. Tirronen, J. Torsner, and Y. Koucheryavy, "Efficient small data access for machine-type communications in LTE," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 3569–3574.
[12] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.
[13] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma, and P. Zhu, "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Global Commun. Conf. Workshop (Globecom Workshop)*, Austin, TX, USA, Dec. 2014, pp. 900–905.
[14] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for Internet of Things," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, May 2016, pp. 1–5.
[15] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—A novel nonorthogonal multiple access for fifth-generation radio networks," *Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, Apr. 2017.
[16] Y. Cao, H. Sun, J. Soriaga, and T. Ji, "Resource spread multiple access—A novel transmission scheme for 5G uplink," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.
[17] 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. 2018.
[18] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, Nov. 2016.
[19] J. Zhang, Y. Pan, and J. Xu, "Compressive sensing for joint user activity and data detection in grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 857–860, Jun. 2019.
[20] S. Dogan, A. Tusha, and H. Arslan, "NOMA with index modulation for uplink URLLC through grant-free access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1249–1257, Oct. 2019.
[21] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23602–23611, 2018.
[22] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.
[23] S. Sesia, I. Toufik, and M. Baker, *LTE-the UMTS long term evolution: From theory to practice*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.
[24] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.
[25] *Evolved Universal Terrestrial Radio Access(E-UTRA); Physical Layer Procedures*, document 3GPP TS 36.213 V12.7.0, Sep. 2015.

TAEHOON KIM (S'13–M'17) received the B.S. degree in media communications engineering from Hanyang University, Seoul, South Korea, in 2011, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute for Science and Technology (KAIST), Daejeon, South Korea, in 2013 and 2017, respectively. He has been with the Agency for Defense Development (ADD), South Korea, as a Senior Researcher, since September 2017. His research interests include wireless communications, 5G, machine-type communications (MTC), the Internet-of-Things (IoT), military communications, physical-layer security, wireless avionics intra-communications (WAIC), and laser detection and ranging (LADAR).

BANG CHUL JUNG (S'02–M'08–SM'14) received the B.S. degree in electronics engineering from Ajou University, Suwon, Korea, in 2002, and the M.S. and Ph.D. degrees in electrical and computer engineering from KAIST, Daejeon, South Korea, in 2004 and 2008, respectively.

From January 2009 to February 2010, he was a Senior Researcher/Research Professor with the KAIST Institute for Information Technology Convergence, Daejeon. From March 2010 to August 2015, he was a Faculty Member of Gyeongsang National University, Tongyeong, South Korea. He is currently a Professor with the Department of Electronics Engineering, Chungnam National University, Daejeon. His research interests include wireless communications, statistical signal processing, information theory, interference management, radar signal processing, spectrum sharing, multiple antennas, multiple access techniques, radio resource management, machine learning, and deep learning. Dr. Jung was the recipient of the 5th IEEE Communication Society Asia–Pacific Outstanding Young Researcher Award, in 2011, KICS Haedong Young Scholar Award, in 2015, and the 29th KOFST Science and Technology Best Paper Award, in 2019. He has been serving as an Associate Editor of *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences, since 2018*.