Adaptive Resource Allocation and Congestion Control Algorithm for Massive Devices in LTE-A

Sung-Hyung Lee, So-Yi Jung, and Jae-Hyun Kim
Department of Electrical and Computer Engineering
Ajou University
Suwon, Korea
xaviersr@ajou.ac.kr, sogloomy@ajou.ac.kr, jkim@ajou.ac.kr

***Abstract*—Internet of things (IoT) devices in the long term evolution-advanced (LTE-A) network require the random access (RA) for the data transmission. The base station in LTE-A requires a decision algorithm for the number of preambles and for the probability of devices to enter contention. This paper proposes an adaptive resource allocation and congestion control algorithm referring to the most recently observed contention results. Furthermore, this paper also proposes an estimation method which estimates the number of contending devices and the number of activated devices in the network based on the unused number of preambles. The performance evaluation using the RA simulator shows that the proposed algorithm can achieve the throughput which is close to the optimal throughput. We also address the limit of current LTE-A system to support massive number of devices based on the evaluation results.**

***Index Terms*—adaptive resource allocation, congestion control, random access, massive devices, Internet of things**

I. Introduction

Due to the variety of objects, Internet of things (IoT) can includes various application areas such as smart grids, smart homes, intelligent transportation, and health care [1]. Because any object can be the IoT device, the number of devices in the network can be massive. 3rd Generation Partnership Project (3GPP) expects 30,000 devices per sector [2], and the 5G Infrastructure Public Private Partnership (5GPPP) expects 100 times more devices than that in 3GPP [3]. A working group in the Radiocommunication Sector of International Telecommunication Union (ITU-R) targets 1 million devices in km^2 [4].

Considering the number of devices, 3GPP is developing their cellular network specifications for long term evolution-advanced (LTE-A) radio technology [5] and new radio (NR) technology [6]. A random access (RA) procedure is used for the data transmission of IoT devices in the network with 3GPP radio technology. The RA procedure requires preamble transmission through the random access channel (RACH), where preambles can be shared with legacy communication service, human-to-human communication service, and IoT service. Therefore, the base stations (BSs) in network such as evolved node B (eNodeB) or gNodeB require an adaptive resource allocation algorithm to allocate the number of preambles for IoT service efficiently rather than sufficiently. In addition, if multiple devices choose same preamble in same RACH, the collision can be happen which causes the backoff of devices. Since the collision probability increases as the number of contending devices in a RACH increases, the BS also requires an congestion control algorithm to limit the number of contending devices per RACH.

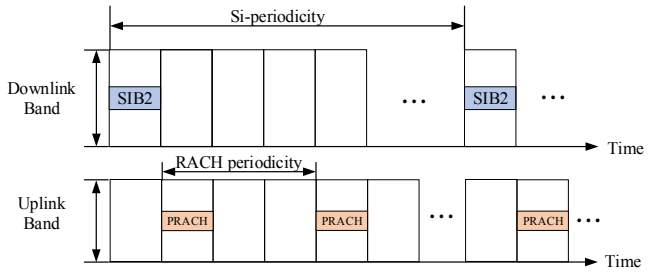
The adaptive resource allocation is studied to achieve maximum throughput. In [7], the number of preambles is adjusted to achieve maximum throughput without limit of the number of preambles. The access class barring (ACB) is also studied where ACB limits the number of contending devices per RACH by using a probability to enter contention. The heuristic algorithm based probability selection algorithm is given in [8], and the maximum likelihood estimator based probability selection algorithm for ACB is given in [9]. However, these studies did not consider both the adaptive resource allocation and ACB simultaneously. Recently, adaptive resource allocation and ACB are simultaneously considered in [1], but the evaluation does not consider the interval to update the number of preambles and the probability to enter contention.

This paper proposes a "most recent (MR)" algorithm which adaptively changes both the number of preambles for IoT services and the probability to enter contention based on the most recent contention result. This paper also proposes the estimation method to estimate the number of contending devices and the number of activated devices which can be used in the LTE-A system. In addition, this paper evaluates the performance of algorithm considering the interval to update the number of preambles and the probability to enter contention.

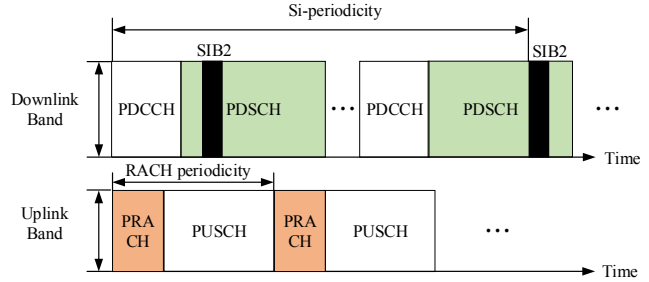
II. Data transmission for IoT devices in LTE-A

A. Frame Structure and Transmission of System Information

Fig. 1(a) shows the frame structure for the IoT devices in wideband LTE-A system (1.4 MHz or above), and Fig. 1(b) shows that in narrow band IoT (NB-IoT) system in LTE-A. In both systems, the frequency bands are divided into downlink band and uplink band. The downlink band includes physical downlink control channel (PDCCH) and physical downlink shared channel (PDSCH) where both channels are used for the transmission of messages. The uplink band includes physical random access channel (PRACH) and physical uplink shared channel (PUSCH), where PRACH is used for the transmission of preambles and PUSCH is used for the transmission of messages. PRACH is allocated periodically by the BS [10], where the periodicity is defined as "RACH periodicity" in



(a)



(b)

Fig. 1. Allocation of PRACH and transmission of SIB in LTE-A (a) wideband LTE (b) NB-IoT

this paper. The BS periodically broadcasts system information block (SIB) using a part of downlink band which includes the information for devices. The SIB type 2 (SIB2) includes information related to RA. The periodicity of the transmission of SIB2 is defined as "si-periodicity" in this paper [10].

B. Data Transmission Procedure

A device is activated when a data to transmit through BS is generated in the upper layer of the device, or it receives a paging message from BS for downlink data transmission. The device first obtains the system information related to RA by receiving SIB2. For the adaptive resource allocation and congestion control of RA, the SIB includes the pool of preambles and the probability to enter contention. The device chooses a random value $y \in [0, 1]$. If the y is less than the probability to enter contention, the device selects a preamble from the preamble pool where the preamble is a code sequence orthogonal to other preambles. Otherwise, it defers RA to next RACH. If the device selected a preamble, it transmits the selected preamble through PRACH. If the BS detected the preamble, it transmits a random access response (RAR, it is also referred as MSG2) to devices which are transmitted the detected preamble. If the device receives the MSG2 corresponding to the transmitted preamble, the device sends third message (MSG3) to the BS, where MSG3 is generally a control message from radio resource control (RRC) layer in the device. If MSG3 is decoded in BS, the BS responds by transmitting contention resolution message (MSG4). Note

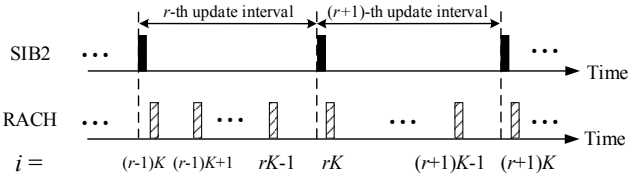


Fig. 2. SIB2 transmission and RACH allocation

that the procedure from the transmission of preamble to the transmission of MSG4 is called as RA procedure in LTE-A.

The data is transmitted from or to the device by a further procedure after the transmission of MSG4. The further procedure includes the transmission of the response in RRC layer, resource allocation by BS for the transmission of data, and data transmission using the allocated resource. The detail in the further procedure are different in conventional LTE, cellular IoT (CIoT), and NB-IoT in LTE-A. If MSG3 is not collided, the messages transmitted in later steps can be delivered to their destination with high probability by the hybrid automatic repeat request (HARQ). For example, 3GPP TR 37.868 [2] assumes 10% loss probability of the single transmission in HARQ, where this ensures 99.9999% of delivery probability for a message using HARQ. Therefore, the successful delivery of MSG3 can be regarded the success of data transmission, where this assumption is general in studies for the random access of LTE-A [1], [7].

III. System Model

Consider a cell with an BS and M of IoT devices in the coverage of the BS. A device activates at t-th subframe where $t \in [0, T_A]$ for uplink data transmission where T_A is a maximum activation time. t is determined by an arbitrary arrival distribution. Let I_A be the RACH interval. Let i be the index of RACH where i is positive integer. i-th RACH is allocated in the iI_A-th subframe. If a device is activated in a t_a-th subframe where $t_a \in [(i-1)I_A, iI_A - 1]$, its first RA trial can be performed in i-th RACH.

Let K be the number of RACHs included in a si-periodicity and r be the index of si-period, respectively. K and r are positive integers. The BS can change the number of preambles per RACH ("pool size" in this paper) and the probability to enter contention ("ACB factor" in this paper) at the start of rK-th RACH. Let r-th update interval include RACHs from $((r-1)K)$-th RACH to $(rK-1)$-th RACH, which means that i-th RACH is included in r-th update interval with the relation of

$$r = \left\lfloor \frac{i-1}{K} \right\rfloor + 1, \tag{1}$$

where $\lfloor x \rfloor$ is the largest integer smaller than x. Fig. 2 shows the relation between i and r.

Let R_r be the pool size and p_r be the ACB factor for r-th update interval, respectively, where R_r and p_r are sent by SIB2 to devices. An activated device generates a random variable

$y \in [0,1]$. If $y > p_r$, the device defers to next RACH and generates y again. Otherwise, the device chooses a preamble from R_r preambles and transmits the selected preamble to BS in RACH. The transmitted preamble is always detectable in BS. If a preamble is detected, the BS transmits an MSG2 after T_{RAR} subframes for the devices those transmitted the detected preamble. If a device receives MSG2 corresponding to their transmitted preamble, it transmits its MSG3 to the BS. If the transmitted preamble is selected by multiple devices in a RACH, the MSG3 experiences collision, and the BS cannot send MSG4. Otherwise, the BS transmits MSG4 to the device as the response of MSG3, and the RA of the device is completed as success. If a device experiences the collision of MSG3, it knows the collision after T_{MSG4} subframes from the time of MSG3 transmission. If the device knows the collision, it does backoff: the device selects an random integer in $[0, W_{BO}]$ where W_{BO} is backoff indicator, and waits multiple subframes where the waiting time is equal to selected random integer, then generates y.

Let M_i be the number of activated devices and N_i be the number of contending devices in i-th RACH, respectively. Let S_i be the number of successful devices in i-th RACH, where success means that the device chooses a preamble which is not selected by other device. Let T_i be the throughput in i-th RACH which is defined as

$$T_i = \frac{S_i}{R_r}. \tag{2}$$

The BS selects two parameters, R_r and p_r in every update interval to optimize throughput, where these parameter changes M_i, N_i, and S_i. In addition, the BS targets to increase S_i if the throughput can be maintained in optimal condition.

IV. Dynamic Resource Allocation and Congestion Control Algorithm

A. Theoretical Optimal Throughput

Our proposed algorithm is based on the parameter selection algorithm given that the BS can know all information and the BS can change the parameters immediately, where these conditions are generally impossible in real system. Let P_i be the probability of the success of a RA in i-th RACH, which is the expectation of the throughput in LTE-A. This probability is equal to the probability that a device selects a preamble where N_i devices select their preambles from R_r preambles, as follows:

$$\mathbb{E}[T_i|N_i, R_r] = P_i = \begin{pmatrix} N_i \\ 1 \end{pmatrix} \frac{1}{R_r} \left(1 - \frac{1}{R_r}\right)^{N_i - 1}, \tag{3}$$

where $\mathbb{E}[\cdot]$ is statistical expectation. The throughput shows maximum in $N_i = R_r$ with $\mathbb{E}[T_i|N_i, R_r] \simeq e^{-1}$, which can be obtained from $d\mathbb{E}[T_i|N_i, R_r]/dN_i = 0$. Since the number of bins is R_r, the expected number of successful devices, $\mathbb{E}[S_i|N_i, R_r]$, is equal to

$$\mathbb{E}[S_i|N_i, R_r] = R_r P_i = N_i \left(1 - \frac{1}{R_r}\right)^{N_i - 1}. \tag{4}$$

When $M_i \leq R_{max}$, S_i is increased as N_i increases where $N_i \leq M_i$. To maximize both S_i and T_i, the BS will select $p_r = 1$ and $R_r = M_i = N_i$. When $M_i > R_{max}$, R_r should be R_{max}, otherwise S_i is decreased. Since R_r is fixed, the BS requires to adjust p_r to satisfy $\mathbb{E}[N_i|M_i, p_r] = R_r$. Therefore, the adaptive resource allocation and congestion control algorithm becomes as follows: the BS first selects $R_r = \min(M_i, R_{max})$ then adjusts $p_r = \min(1, M_i/R_r)$ to optimize throughput and to maximize S_i.

B. Proposed "Most Recent (MR)" Algorithm for LTE-A Network

In actual system, the BS cannot obtain or estimate both M_i and N_i until i-th RACH is completed. Fortunately, the estimation of M_{i-1} can be available in the BS before the start of i-th RACH. Furthermore, M_{i-1} and M_i have correlation due to the backoff of devices. Thus, the BS can decide parameters based on M_{i-1} instead of M_i. Let $\hat{M}_{i-1}$ be the estimation of M_{i-1}. Replacing M_i in the adaptive resource from the theoretical throughput optimization to $\hat{M}_{i-1}$, we can obtain following MR algorithm: the BS first selects $R_r = \min(\hat{M}_{i-1}, R_{max})$ then adjusts $p_r = \min(1, \hat{M}_{i-1}/R_r)$.

C. Proposed Estimation Algorithm of the Number of Devices for LTE-A Network

For the proposed most recent algorithm, $\hat{M}_{i-1}$ is required in BS. In the LTE-A, the success or collision is decided after reception of MSG3, which requires the waiting of multiple subframes more than RACH periodicity. On the other hand, the number of unused preambles can be obtained immediately after completion of a RACH. Let $O_{0,i}$ be the observed number of unused preambles in i-th RACH. Suppose that N_{i-1} devices are competing using R_x preambles in $(i-1)$-th RACH. The expectation for the observed number of unused preambles in $(i-1)$-th RACH is equal to

$$\mathbb{E}[O_{0,(i-1)}] = R_x \left(1 - \frac{1}{R_x}\right)^{N_{i-1}}. \tag{5}$$

Rearranging (5) for N_{i-1} results following equation:

$$N_{i-1} = \frac{\log\left(\frac{\mathbb{E}[O_{0,(i-1)}]}{R_x}\right)}{\log\left(1 - \frac{1}{R_x}\right)}. \tag{6}$$

In actual system, the BS can obtain $O_{0,(i-1)}$ but not N_{i-1}. Let $\tilde{N}_{1,i-1}$ be the estimation of N_{i-1} based on (6). From (6), the BS can get $\tilde{N}_{1,i-1}$ using $O_{0,(i-1)}$ as follow:

$$\tilde{N}_{1,i-1} = \frac{\log\left(\frac{O_{0,(i-1)}}{R_x}\right)}{\log\left(1 - \frac{1}{R_x}\right)} \tag{7}$$

$\tilde{N}_{1,i-1}$ becomes infinity when $O_{0,(i-1)} = 0$ which is not useful for BS.

For $O_{0,(i-1)} = 0$, the BS can consider that N_{i-1} is large enough to use all preambles. Let $\tilde{N}_{2,i-1}$ be the another

estimation of N_{i-1} for $O_{0,(i-1)} = 0$. We propose the another estimation as follows especially for $O_{0,(i-1)} = 0$:

$$\tilde{N}_{2,i-1} = 2(R_x - O_{0,(i-1)}). \tag{8}$$

Let $\hat{N}_{i-1}$ be the finalized estimation of N_{i-1}. The BS with proposed algorithm obtains $\hat{N}_{i-1}$ as following equation:

$$\hat{N}_{i-1} = \begin{cases} \tilde{N}_{1,i-1} = \frac{\log\left(\frac{O_{0,(i-1)}}{R_x}\right)}{\log\left(1-\frac{1}{R_x}\right)}, & ;(O_{0,(i-1)} > 0), \\ \tilde{N}_{2,i-1} = 2R_x & ;(O_{0,(i-1)} = 0). \end{cases} \tag{9}$$

Suppose that p_x is used as the ACB factor in $(i-1)$-th RACH. The statistical expectation of N_{i-1} for given M_{i-1} and p_x is equal to

$$\mathbb{E}[N_{i-1}|M_{i-1}, p_x] = M_{i-1}p_x. \tag{10}$$

Based on (10), M_{i-1} can be estimated by replacing $\mathbb{E}[N_{i-1}|M_{i-1}, p_x]$ to $\hat{N}_{i-1}$ and by arranging the equation as follows:

$$\hat{M}_{i-1} = \frac{\hat{N}_{i-1}}{p_x}. \tag{11}$$

The BS with proposed estimation method acts as follows to obtain $\hat{M}_{i-1}$: The BS obtains $\hat{N}_{i-1}$ using (9) and then obtains $\hat{M}_{i-1}$ using (11).

V. Performance Evaluation

A. Evaluation Method and Setup

The LTE-A RA simulator using the Riverbed Modeler (known as OPNET) is used for evaluation. In the simulator, an BS node and multiple number of devices are deployed for the evaluation of the proposed most recent algorithm. The arrival time distribution for devices is set as uniform or Beta ($\alpha = 3$, $\beta = 4$) distribution [2]. T_A is selected as 10,000 subframes [2]. I_A is selected as 5 subframes [2]. T_{RAR}, T_{MSG4}, and W_{BO} are selected as 5, 48, and 20 subframes, respectively [2]. R_{min} is selected as 3 to avoid high estimation error in low traffic load [7]. R_{max} is selected as 64 [11] and the initial number of preambles (R_1) is selected as 54 [2].

B. Number of contending devices and ACB factor

Fig. 3(a) shows the number of contending devices (N_i) and Fig. 3(b) shows the ACB factor (p_r) chosen by BS in time, respectively. In this figure, M is 50,000, K is 1, and devices are arrived with uniform distribution. Note that, pool size are maximum in this scenario. If the BS can change two parameters, R_r and p_r, in every completion of RACH, N_i changes around a certain mean number of contending devices as shown in Fig. 3 (a).

Fig. 4 shows N_i and p_r for same case but K is 32. Note that, pool size are also maximum in this scenario. If the BS cannot change the parameters in every RACH as in Fig. 4 (b), N_i can be changed to a value far from a certain average number of contending devices. For example, N_i suddenly increases in first K RACHs due to lack of preambles and high ACB factor, but the BS cannot change p_r and R_r until K RACHs are passed. We can expect that the sudden increase or decrease of N_i will decrease throughput for large K since the throughput is maximized when $N_i = R_r$ and it decreases as $|N_i - R_r|$ increases.

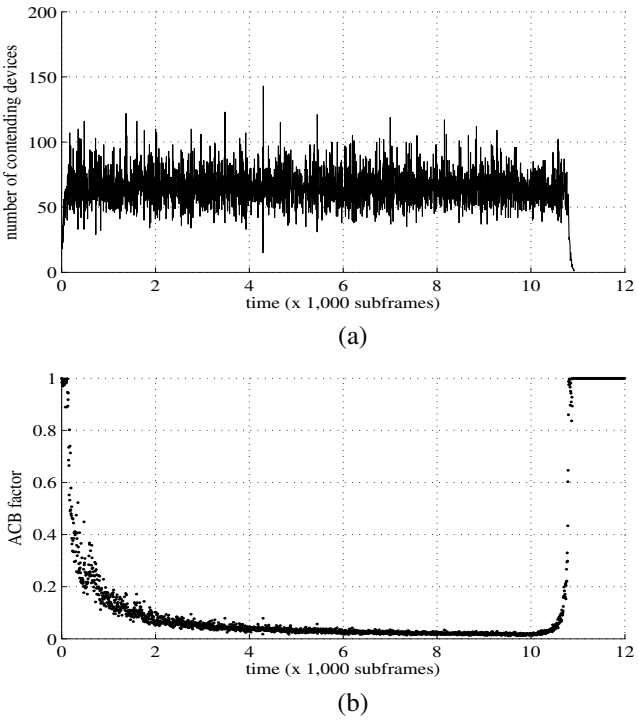


Fig. 3. (a) Number of contending devices (N_i) in time and (b) ACB factor (p_r) in time when K=1

C. Averaged Throughput and Delay

For the averaged throughput and delay, we did 1,000 simulations per each point. The throughput and delay are collected until all devices succeed their RA. The standard error for each point is less than 0.5% of averaged value.

Fig. 5 shows the average throughput with uniformly distributed arrivals. The throughput of proposed algorithm is compared with that of dynamic ACB with dynamic resource allocation (D-ACB with DRA) in [1] with a parameter $b = 1.0$. Note that, D-ACB with DRA requires following assumption in addition to our system model: The number of successful and collided preambles can be known in the BS when the BS receives preamble. The proposed algorithm shows high throughput which is close to the maximum achievable throughput ($\simeq 0.36$). The throughput slightly decreases when $K = 32$ due to the increase of $|N_i - R_r|$ as shown in Fig. 4. The throughput in low traffic load is lower than that in high traffic load, where low traffic load means small number of deployed devices in cell. In low traffic load situation, the ACB factor becomes 1 and the BS changes the pool size to maximize throughput. Since the pool size can only be integer, thus the precision of algorithm decreases which also decreases average throughput. D-ACB with DRA shows lower throughput than proposed algorithm in low or medium traffic load because D-

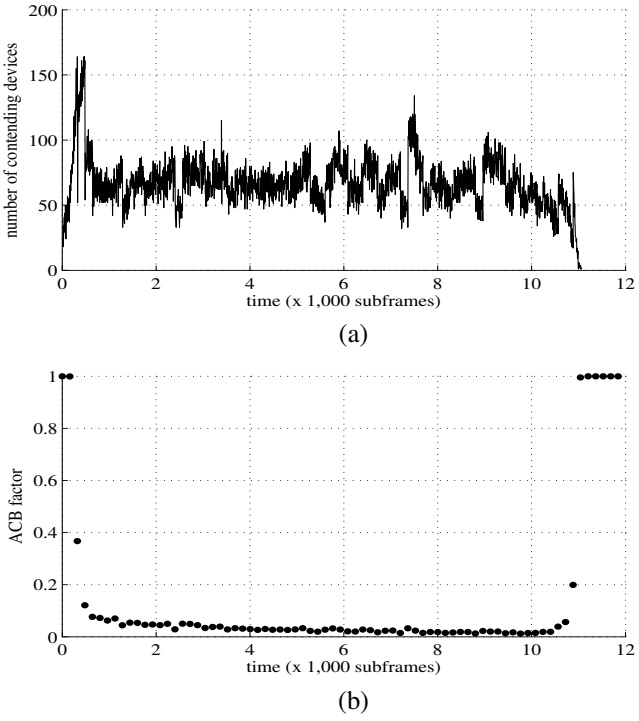


Fig. 4. (a) Number of contending devices (N_i) in time and (b) ACB factor (p_r) in time when K=32

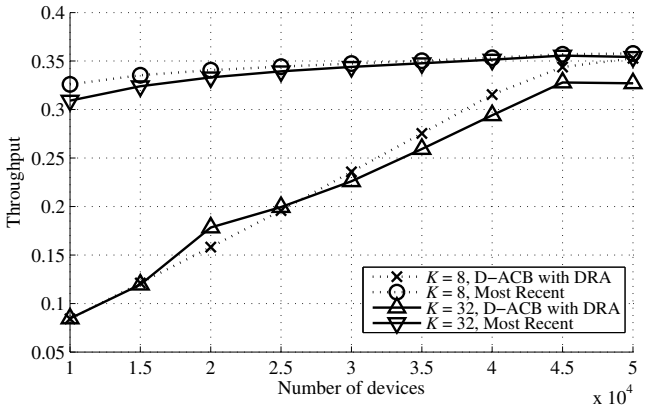


Fig. 5. Average throughput vs. number of deployed devices with uniform distributed arrival

ACB with DRA selects high pool size in low traffic load due to different objective.

Fig. 6 shows average delay for uniformly distributed arrival. When the number of preambles is sufficient to serve all activated devices, the delay decreases as throughput increases. However, if the maximum number of preamble is not sufficient as the delay with 50,000 devices, the delay increases since ACB limits the number of contending devices per RACH. The average delay of D-ACB with DRA for $K = 8$ generally shows lower delay than proposed algorithm since the number of successful devices per RACH is larger than that in proposed algorithm, where large number of successful devices is from

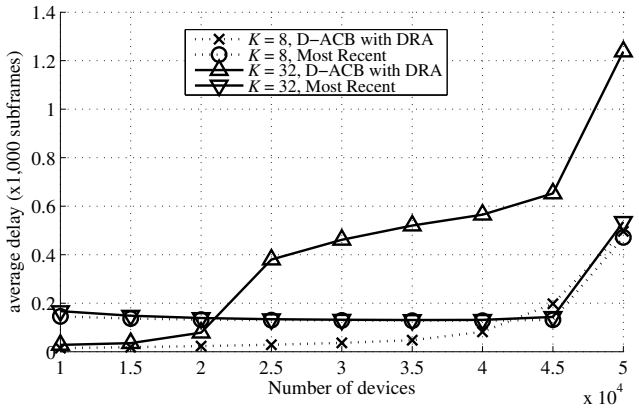


Fig. 6. Average throughput vs. number of deployed devices with uniform distributed arrival

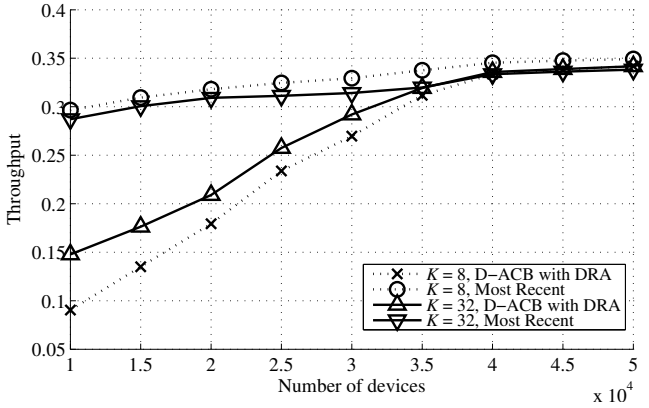


Fig. 7. Average throughput vs. number of deployed devices with Beta distributed arrival

large number of assigned preambles. D-ACB with DRA with $K = 32$ shows higher delay above 25,000 devices than other cases. In this case, the BS periodically selects very low pool size with low ACB factor which unnecessarily defers RA of devices.

Fig. 7 and Fig. 8 shows the throughput and delay, respectively, for Beta distributed arrival. The Beta distributed arrival shows $\cap$-shaped arrival rate in $t \in [0, T_A]$, and the changes of arrival rate causes the rapid change in the number of contending devices. In the Beta distributed arrival, the proposed algorithm also shows higher throughput than D-ACB with DRA. The average delay of proposed algorithm shows similar to that of D-ACB with DRA.

Fig. 9 shows average throughput with respect to K, and Fig. 10 shows average delay. The system shows better throughput and delay with small K because the BS can change the status of network frequently as K decreases. Therefore, we recommend that the BS chooses small K for better throughput and delay performance, where the frequent transmission of SIB2 will require more resources in downlink band.

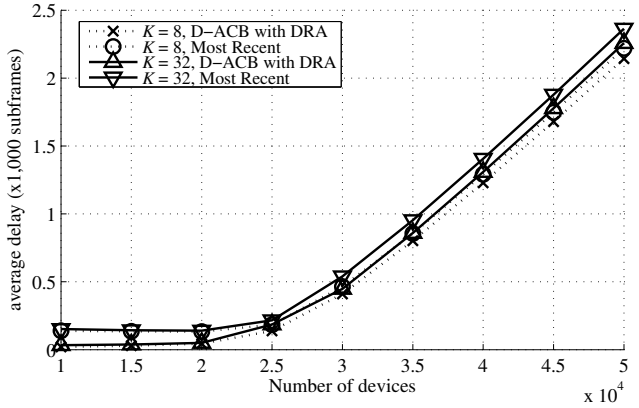


Fig. 8. Average delay vs. number of deployed devices with Beta distributed arrival

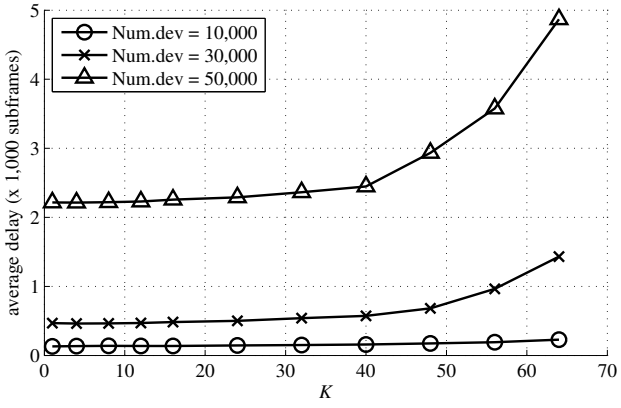


Fig. 10. Average delay vs. K with Beta distributed arrival for proposed algorithm

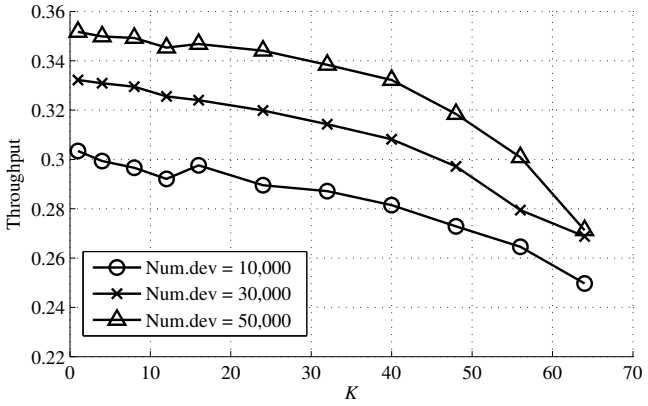


Fig. 9. Average throughput vs. K with Beta distributed arrival for proposed algorithm

D. Discussion

The performance evaluation results show that the BS requires to select K as 1 to maximize throughput. For $K > 1$, the throughput degradation is hard to avoid in current LTE-A since SIB2 includes single pool size and single ACB factor. In order to increase the throughput while reducing the resource consumption, the SIB2 may include the sequence of pool size and the sequence of ACB factor where each element in the sequence can be used for each corresponding upcoming RACH. The selection of the sequence of pool size and the sequence of ACB factor can be the future work for LTE-A system to increase throughput for the cases with K larger than 1.

VI. Conclusion

This paper reviewed the RA procedure in LTE-A for massive IoT devices and proposed a MR algorithm, which is an adaptive resource allocation and congestion control algorithm for better throughput. The proposed algorithm first changes the pool size to maximize the throughput while increases the number of successful devices, then selects ACB factor to maximize throughput for changed pool size. The proposed algorithm also includes the estimation algorithm for BS to obtain the number of contending devices and the number of activated devices considering the limit in the LTE-A system. The performance results show that the proposed algorithm can be used to achieve good throughput and delay in the LTE-A system. In addition, we expect that the proposed algorithm can be used for new radio (NR) technology in 3GPP since the RA procedure in NR is very similar to that in LTE-A.

References

[1] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9847–9861, December 2016.

[2] 3GPP TR 37.868 v11.0.0, "RAN Improvements for Machine-type Communications," October 2011.

[3] ICT-317669-METIS/D1.1, "Scenarios, requirements and KPIs for 5G mobile and wireless system," April 2013.

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

[5] 3GPP TR 36.321 v14.3.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (Release 14)," June 2017.

[6] 3GPP TR 38.321 v0.2.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; New Radio; Medium Access Control (MAC) protocol specification (Release 15)," July 2017.

[7] J. Choi, "On the Adaptive Determination of the Number of Preambles in RACH for MTC," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1385–1388, July 2016.

[8] S. Duan, V. Shah-Mansouri, and V. W. S. Wong, "Dynamic access class barring for M2M communications in LTE networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, December 2013, pp. 4747–4752.

[9] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5815–5820.

[10] 3GPP TR 36.331 v13.2.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 14)," March 2017.

[11] 3GPP TR 36.211, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 14)," March 2017.