Dynamic Resource Reservation for Ultra-low Latency IoT Air-Interface Slice

Guolin Sun[1], Guohui Wang[1], Prince Clement Addo[1], Guisong Liu[1], Wei Jiang[2,3]
[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu, Sichuan 611731 – P. R. China
[e-mail: guolin.sun@uestc.edu.cn]
[2] German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany
[3] Department of Electrical and Information Technology (EIT), Technische University (TU) Kaiserslautern, Germany
[e-mail: wei.jiang@dfki.de]
*Corresponding author: Guolin Sun

Received May 29, 2016; revised September 30, 2016; accepted October 30, 2016; published July 31, 2017

Abstract

The application of Internet of Things (IoT) in the next generation cellular networks imposes a new characteristic on the data traffic, where a massive number of small packets need to be transmitted. In addition, some emerging IoT-based emergency services require a real-time data delivery within a few milliseconds, referring to as ultra-low latency transmission. However, current techniques cannot provide such a low latency in combination with a mice-flow traffic. In this paper, we propose a dynamic resource reservation schema based on an air-interface slicing scheme in the context of a massive number of sensors with emergency flows. The proposed schema can achieve an air-interface latency of a few milliseconds by means of allowing emergency flows to be transported through a dedicated radio connection with guaranteed network resources. In order to schedule the delay-sensitive flows immediately, dynamic resource updating, silence-probability based collision avoidance, and window-based re-transmission are introduced to combine with the frame-slotted Aloha protocol. To evaluate performance of the proposed schema, a probabilistic model is provided to derive the analytical results, which are compared with the numerical results from Monte-Carlo simulations.

Keywords: ultra-low latency transmission; network slice; resource reservation

1. Introduction

Achieving full automation to enhance the sensory and processing capabilities of human beings has been regarded as one of the ultimate goals of the tactile Internet. The goal of Ultra-Low Latency (ULL) embraces all the upcoming applications such as unmanned or remote control, augmented reality, intelligent transportation systems, smart grid and the Internet of Things (IoT). There are three major characteristics that distinguish it from conventional services such as multimedia distribution and file delivery.

First, remote control, data collection and environmental sensing play a considerable role in the ULL transmission. These applications generate “mice” flows with few coded bytes in each small packet [1]. As the number of IoT sensors involved in these applications is tremendous, the amount of such small packets can be extremely large. To support such massive number of intermittent small-packet transmissions, the complexity of radio resource allocation, scheduling and routing information exchanging becomes challenging.

Second, these applications are vulnerable to the end-to-end latency. The state-of-the-art Fourth Generation (4G) mobile networks cannot support ULL transmissions, because heavy signaling overheads have already been imposed on both wired and wireless links. Providing ultra-low-latency services in a few milliseconds consequently turns out to be one of the most crucial requirements in the Fifth Generation (5G) systems. To achieve this requirement, the latency of the existing mobile networks needs to be reduced 100- to 1000-folds. Obviously, new designs and fundamental inspirations are required.

Third, the traffic pattern of emergency IoT flows is disruptive where most of terminals are in idle states. As we know, the signaling overhead of terminals in the process of Radio Resource Control (RRC) re-establishment in the 4G system is already a burden. When a massive number of IoT sensors switch frequently between an idle mode and a connected mode, the network may be overwhelmed by a huge number of control signals.

The dynamic air-interface slicing is proposed to guarantee various latency levels in order to meet new challenges of ULL applications and massive IoT connections. This approach is optimized for the virtualized resource sharing in the air interface and is perfectly applicable in mobile networks due to limited bandwidth resource [2]. In this paper, we propose a dynamic resource reservation method to mitigate the problem of large signaling overhead for disruptive emergency flows. ULL flows can temporarily borrow the data plane resource from other slices, e.g., a mobile-UE slice. The conventional resource reservation always statically allocates resources for data transmission, whereas the dynamic resource reservation works in a different way. First, it is a persistent method, rather than an overlaying on “unused” resource blocks (RBs). Second, the reserved resource is virtual and dynamic. It is dynamically mapped onto the physical RBs but temporarily scheduled out from other air-interface slices. The effective bandwidth is computed in advance by a negotiation. Third, it is content centric. Because IoT is a network filled with content-labeled flows, which are identified with names, each flow is treated differently. Finally, our proposed reservation is compatible with LTE frame structure by extending the amount of sub-frames in each frame.

The main contribution of this paper is dedicated to reduce the control signaling latency by resource provisioning when an emergency IoT terminal switches from an idle to a connected mode. The main problems will be how to reserve the RBs and how to 'borrow' the resources from non-ULL slices, in the case that the collision of flows from a content grouping happens. A dynamic retry-window based frame-slotted Aloha protocol and a silence-probability based collision avoidance algorithm are designed for the request transmission of ultra-low latency

flows. The data transmission behind the request is then scheduled by 'borrowing' resource blocks on other non-ULL slices.

The organization of this paper is defined as follows. Section 2 summarizes the related work on advanced resource reservation and ultra-low latency radio resource allocation for the Device-to-Device (D2D) communications. In the section 3, we present the proposed signaling approach based on dynamic resource reservation for RRC re-establishment, as well as a probability model to analyze the achieved performance. The comprehensive performance evaluations are illustrated in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

The resource reservation schemas have been extensively investigated in the area of Quality-of-Experience (QoE) guarantee in wireless and wired networks [3]-[6]. The advanced reservation ranges from immediate to future reservation. Immediate reservation can be viewed as starting at “now” and future reservation is the advance reservation of resources for some time point in the future. Mingju and Adachi propose immediate resource reservation schemes for mobile users to achieve fast handover failure recovery and an efficient resource utilization in air-interface [3][4]. Shehada and Liu present some future reservation approaches to the Quality-of-Service (QoS) guaranteed video transmission in wireless network [5][6]. Reserved Resources (RR) in the different systems may be link bandwidth, frame, resource block or buffer space. Except for the radio resource reservation, link bandwidth reservation is proposed in the 3rd Generation Partnership Project (3GPP) Evolved Packet Core (EPC) network [7][8]. This is to pre-allocate the bearer information from application side of network, and reduce the signaling each time when the device wants to communicate. In addition to resource reservation, the QoS/QoE guarantee in wireless and wired networks can also be realized with software-defined networking type of agile network control. An optimization framework is proposed for the OpenFlow controller in order to provide QoS support for scalable video streaming over an OpenFlow network [9][10]. The authors proposed several multi-path provisioning algorithms for cloud-assisted scalable video coding streaming in heterogeneous networks [11]. Authors in [12] solved the problem of the multi-source multi-destination scalable coding video multicast in the OpenFlow controlled network. The authors design several algorithms to allocate bandwidth intelligently and ensure high-quality video streaming.

In recent D2D applications, a resource reservation schema with channel quality detection is recommended [9]. This schema is in a frequency-time grid in terms of RBs in LTE system, as opposed to classical reservation Aloha where time slots are reserved over the whole bandwidth. Regarding resource allocation in D2D communications based on LTE, 3GPP has already started the standardization for D2D communication in Release 12, towards the public safety usage. Till now, most of related works focus on the under-laying D2D communications in order to maximize the spectrum utilization of cellular networks. Besides, data collision in the distributed random access has a critical impact on the ULL services. In [13], a random access protocol for collision avoidance was proposed by introducing the extra response step after random access. Unfortunately, there is no any feedback link in the data plane considered in the LTE D2D communication.

Recently, a collision-aware resource reservation is proposed for the D2D communications [14]. The basic idea is to utilize the unused resource in data region for possible distributed coordination to avoid the collisions. The scheduling assignment region in a frame is used to reserve the data resource for the next scheduling frame. Although the scheduling assignment

region is intermittent with the unused resource and can mitigate collision, it cannot guarantee an ultra-low latency in a few milliseconds. In [15], a collision resolution algorithm based on splitting trees is proposed on the LTE Physical Random Access Channel (PRACH) to solve the problem of synchronous traffic arriving. However, PRACH resource appears once in each five milliseconds but multiple interactions are required to resolute collisions with thousands of devices activated simultaneously. In [16], an analysis and evaluation of dynamic frame-slotted Aloha is provided for the energy harvesting D2D communications. However, the variable frame length design is not compatible with the LTE system. In [17], an initial analysis on ULL radio access problem for remote control was provided. By reducing Transmission Time Interval (TTI) from 1ms to 100us without retransmission in the air interface, it realized a latency of one millisecond.

3. THE PROPOSED SCHEME

3.1 The proposed scheme

Resource reservation: To mitigate the signaling overhead when a terminal switches from an idle to a connected mode, persistent resource reservation is adopted in our design. Based on the analysis of traffic characteristics of IoT and D2D services, we noticed that the packets are small and all of flows are mice. In addition, IoT applications often can be taken as a content centric "database". Therefore, we considered a content grouping based persistent resource reservation for ultra-low latency IoT flows. Considering the compatibility, the proposed approach can be applicable to LTE eNodeB architecture and Software Defined Network (SDN) based architecture for 5G. In the negotiation procedure, the ultra-low latency application registers on an eNodeB or an SDN controller. The negotiation message includes the content name and a set of traffic parameters to identify the flow. According to the theory of Network Calculus and a queue theory with the traffic characteristics analysis, a flow can be described by the traffic specification T-SPEC (p, k, r, b). It is realized with a leaky bucket traffic shaper. The p, k, r, b represent the peak rate, the maximum packet size, the sustainable rate, and the maximum burst size for a flow, respectively. The definition in [18] tells us that the effective bandwidth relies on both the transmit rates and delay requirements. This is known in advance from the registered ultra-low latency flows by the controller or eNodeB. The effective bandwidth is expressed as:

$$e_D = max\left\{\frac{k}{D}, r, p\left(1 - \frac{D-k/p}{(b-k)/p-r+D}\right)\right\} \quad (1)$$

Intuitively, the effective bandwidth sometimes depends on the transmission rate r, sometimes depends on k over D.

Air-interface resource slicing: The bandwidth of air-interface in 5G can be classified into two slices: ultra-low latency IoT slice and mobile-UE slice. The IoT slice has been defined, which is composed of virtual resource blocks. The RRs in the Fig. 1 are the reserved resources for the ULL IoT flows which are scattered on the air-interface slices. This resource indicated in ultra-low latency IoT slice is sent to UEs in a broadcast sub-frame, which is conFig.d as the first sub-frame in each retry window. This resource indication includes the slot serial number and the RB index. Each terminal will resolute the received bitmap information. Once a terminal has the registered ultra-low latency data to send, it will be transmitted on the reserved RBs and does not need to apply with acknowledgement.

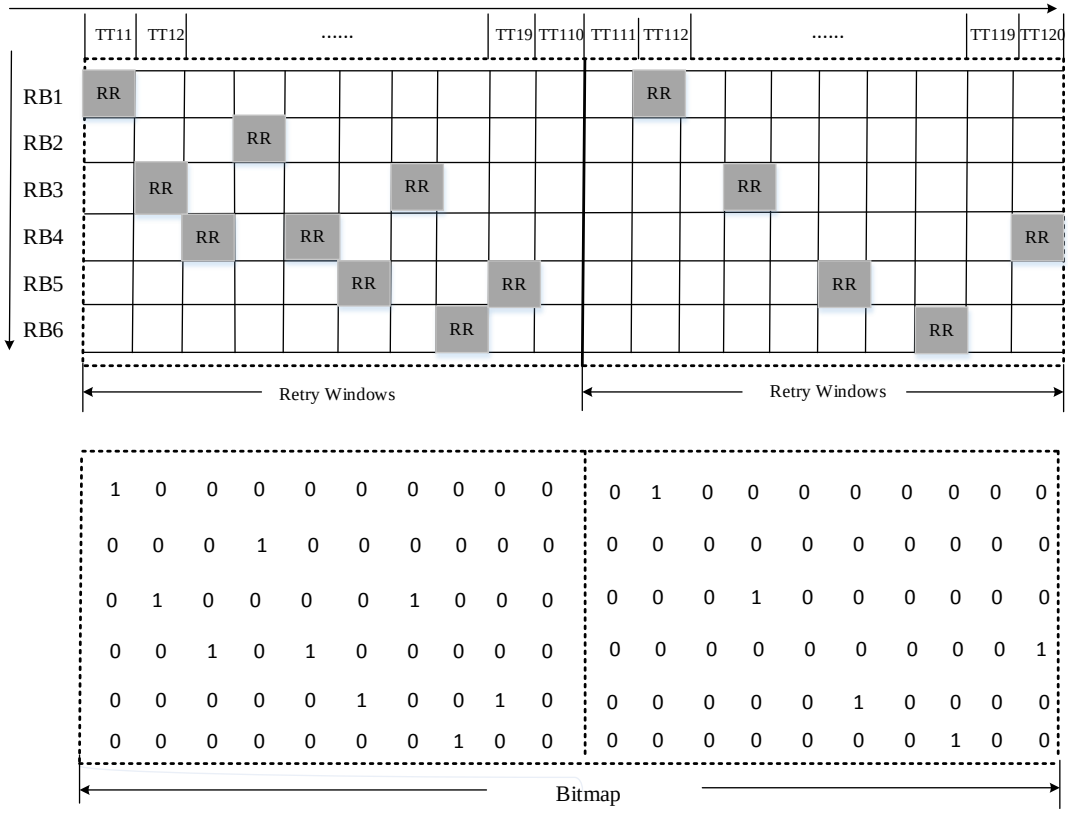

Fig. 1. The Bit-map design for the reserved RBs in the ultra-low latency IoT slice

Fig. 1 illustrates how to reserve RBs for ultra-low latency user access. The bit-map illustrated in **Fig. 1** actually can be changed dynamically in each retry window, based on the collision detection and population prediction of terminals. The LTE network divides a frame into 10 sub-frames with 1ms duration. However, Wi-Fi network divides a frame into many transmission opportunities. The frame structure allows us to estimate the population of contention users in accordance with the Wi-Fi scenario**.** In [19], an optimal contention window estimation method is proposed for Wi-Fi networks. To estimate the population of contention stations, a novel scheme is proposed based on the assumption that the arriving of terminals follows a Poisson distribution.

The optimal window size is defined as

$$RRN_{opt} = \left(N_c + N_q\right) * \sqrt{T_{coll}/2} \tag{2}$$

where, N_c is the number of collisions in the current retry-window, recorded at the eNodeB. N_c is calculated as:

$$N_c = \varepsilon * L \tag{3}$$

and

$$\varepsilon = \frac{\lambda\left(e^{\lambda}-1\right)}{e^{\lambda}-1-\lambda} \tag{4}$$

where, ε is the expected number of contention stations, L is the retry window size, and λ is the mean of Poisson distribution related to the specific scenario. T_{coll} is the average duration of packet collisions in term of TTI in our case. The number of success terminals in each transmission is N_s:

$$N_s = \sum_{i=1}^{L} N_s(i) \tag{5}$$

where, $N_s(i)$ is the number of success terminals in the ith slot, N_s is the sum of $N_s(i)$ in L slots, and N_c is an estimation with equations above. We assume that N_q is the number of silent terminals in the last retry window. Therefore

$$N_q = (N_c + N_s) * P_q / (1 - P_q) \tag{6}$$

In our case, the aggregated flow for a content group shares a common bit-map based RR in ultra-low latency IoT slice, as shown in the **Fig. 1.** It is defined with a bitmap in a frame divided in multiple retry windows. **Fig. 2** illustrates a whole procedure of the proposed protocol. The flows occupy the RRs based on a frame-slotted Aloha in RR region. However, a collision may happen in this RR region. If a collision does not happen as shown in the **Fig. 2**, Flow 1 and Flow 2 will be scheduled in normal data region with resource allocation in the PDCCH. This dynamic resource blocks on two independent slices are illustrated in the **Fig. 2.** It explains the specific behavior in signaling procedures.

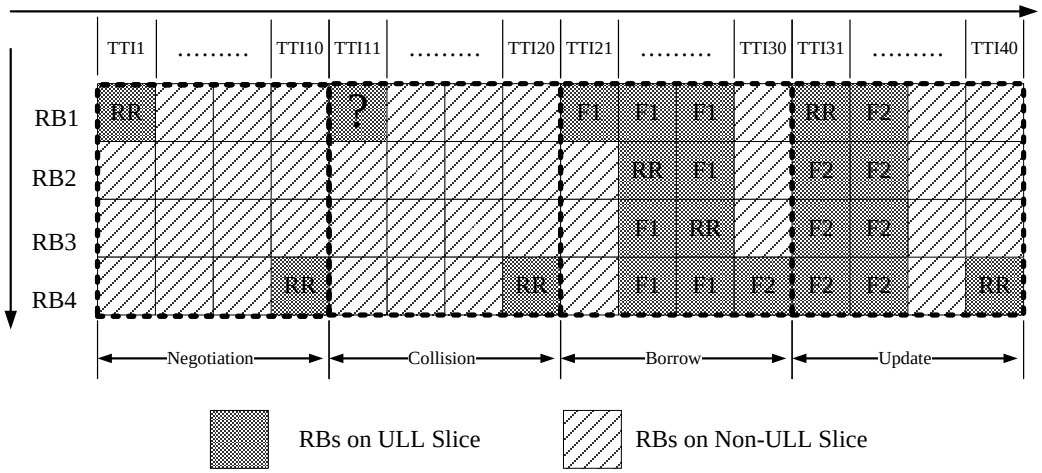

Fig. 2. The sliced resource view on the air-interface

The whole procedure for our proposed approach can be summarized in four steps. First, a user negotiates with the base station or controller whether it can utilize the reserved resource to send its flow. In the **Fig. 2**, it is assumed that initially there are two RBs reserved for each window on the ultra-low latency slice. Second, the registered ultra-low latency flows can access the reserved RBs directly once it has data to transmit. However, if a collision happens on any reserved RBs, they will retry in the next window. We assume a collision happened between Flow 1 and Flow 2 in the **Fig. 2**. In such case there is a retry in the next window. Third, once a base station receives a request from a reserved RB, the RBs on the ULL slice will immediately be extended by borrowing the resources from Non-ULL slice.

This is shown in the case of F1 and F2 in the **Fig. 2**. Finally, the number of reserved RBs in each window is updated dynamically based on collision detection in the previous window. These RBs are always reserved for the incoming ULL flows only. For example, in the **Fig. 2,** the number of reserved RBs is assumed as 2, 2, 4 and 2 for each window respectively.

Data contention: **Fig. 3** illustrates the proposed collision resolution algorithm based on the Frame-slotted Aloha. **Fig. 3** illustrates the collision avoidance in the user access procedure for mobile user when it changes from the idle mode to the connected mode with ULL flows. Therefore, the packet in **Fig. 3** is the first packet in such flow. If no collision happens, the first packet will be detected by eNodeB. Then the eNodeB will borrow the RBs from non-ULLL slice to notify the resource indication to this user in PDCCH. The user will transmit all of remaining packets in its buffer once it gets Acknowledge of the first packet from eNodeB. How many Non-ULL slices should be borrowed depends on the effective bandwidth estimation *eB*, as shown in the equation (1).

We assume that one frame is divided into *M* retry windows and collisions only happen inside of a content group. When an idle terminal has emergency data to transmit, it chooses one of the *M* sub-frames randomly. This random decision is based on the indication in the Master Information Block (MIB). The MIB embraces the reserved RB indication and the value of silence probability *Pq*. In order to mitigate collisions in the RR region in a frame, silence probability is inserted into the MIB message, which appears in every scheduling frame period. This silence probability is conFig.d by the SDN controller based on the traffic load analysis. If a random value generated by this terminal is larger than the predetermined threshold, it keeps "silent" and retries this packet in the next window. Otherwise, it transmits it in one of the *M* sub-frames. Once more than one packet is transmitted in a common RR region, a collision happens. These terminals will re-transmit it again in the next retry window within the same frame. When the number of retries reaches a maximum value, this packet will be discarded.

Extension to Multiple groups: The approach can be extended to multiple groups in two ways: the dedicated and shared resource reservation. Regarding the dedicated, each group occupies an independent RR in term of RBs. In the shared manner, multiple content groups will share a common reserved RB in a random manner. In this case, a collision may happen for the first-packets of many flows from different content groups, but the resource utilization may be improved when the terminals are rare. Moreover, the extension to multiple groups will create more non-ULL slice resource borrowing. The solution needs an agile scheduling scheme to enable ULL flows borrow and compete for the limited Non-ULL slices, which is addressed in another paper[20]. If the shared manner is used to solve the above problem, the amount of shared resources allocated depends on the effective bandwidth estimation *eB* of the grouped flows.

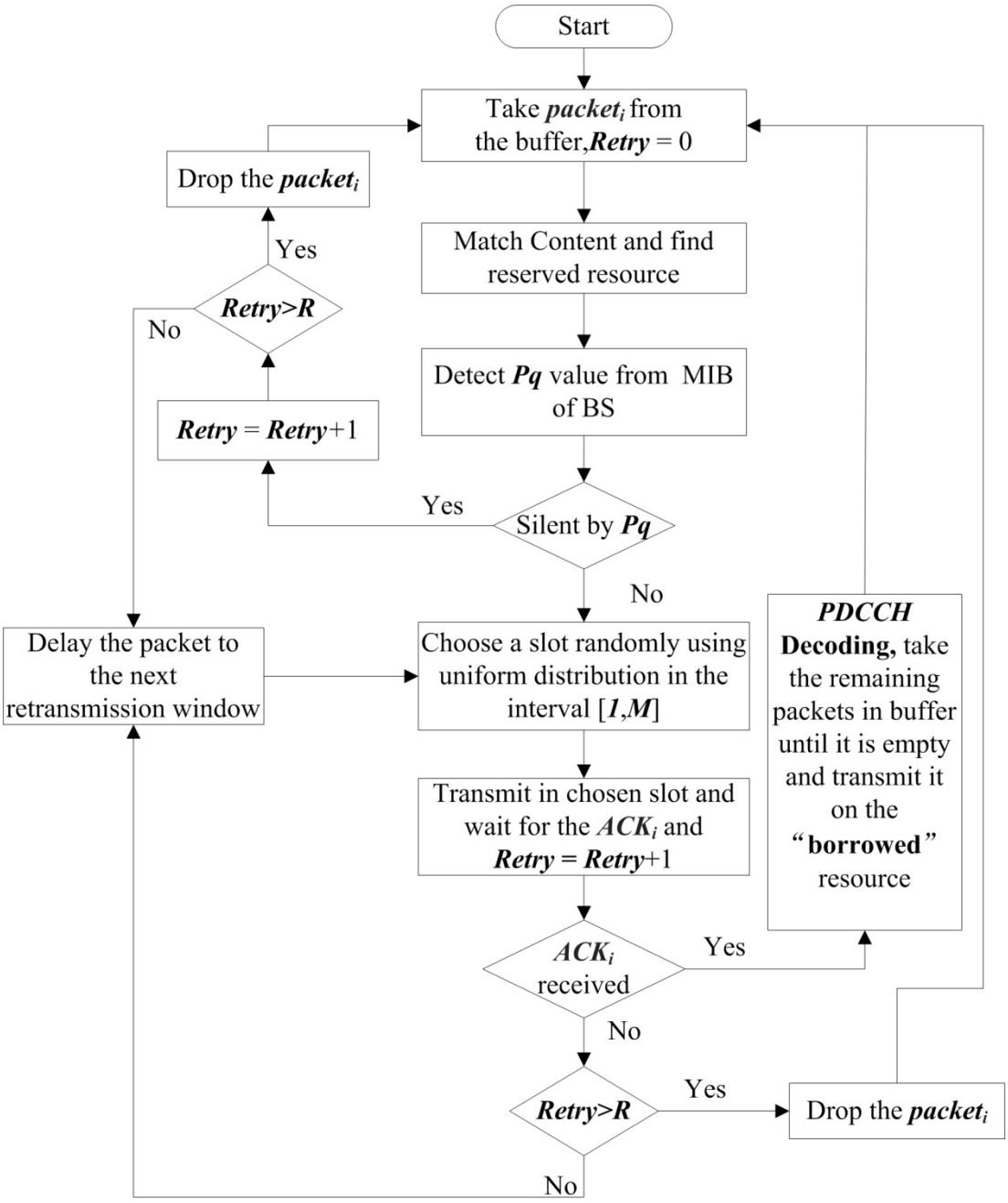

Fig. 3. The proposed collision avoidance algorithm

3.2 Analytic Model

Definition 1: The success probability

This occurs if a terminal chooses one of *M* sub-frames to transmit, and the others with data to transmit choose the remaining (*M-1*) sub-frames. In this scenario, the terminal can successfully transmit. However, the number of terminals is reduced as *N**(1-*Pq*), once the silence frame is executed. Therefore, the success probability for each terminal P_S and the

failure probability P_f are computed as

$$P_s = (1 - P_q) * \left(1 - \frac{1}{M}\right)^{N*(1-P_q)-1} \tag{7}$$

$$P_f = 1 - P_s \tag{8}$$

Definition 2: The drop rate

Once the number of re-transmission times reach the maximum limit for one packet, this packet will be dropped. This happens because collision or silence happens for *R* times. Therefore the drop rate P_d is given as follow [21].

$$P_d = {P_f}^R \tag{9}$$

Definition 3: The delay expectation

The delay expectation is defined as the average waiting time in buffer for a packet. The analytic model is illustrated in the **Fig. 4**. For a terminal, there are *R* opportunities to succeed the transmission in the 1st to the *R*th packets. The probability of success to transmit the packet in the *i*th packet is $P_s(i)$. It is assumed that the failure of transmission happens for *i*-1 times before a packet is dropped. The $P_s(i)$ can be calculated as

$$P_s(i) = {P_f}^{i-1} * P_s \tag{10}$$

We assume that *T* is the duration for each window. Because it takes the total time *i*T* to succeed in transmission at the *i*th packet, the delay expectation for one packet *Ed* is:

$$E_d = \sum_{i=1}^{R} P_s(i) * \left((i-1) * T + T/2\right) + P_d * (R * T) \tag{11}$$

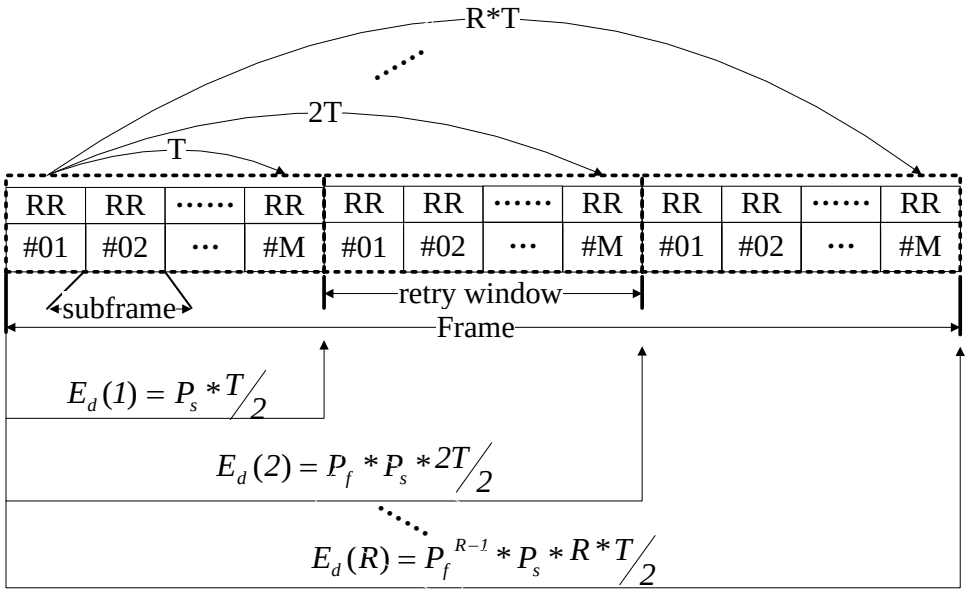

Fig. 4. The analytic model

4. Experimental Results and Analysis

4.1 The testing scenario

Fig. 5 illustrates an instance of simulation scenarios to evaluate the proposed schemes. In our simulation, one eNodeB is deployed with many groups of sensors and mobile terminals. The sensors in Group 2 and Group 3 generate ultra-low latency flows in a random manner, which occupied the reserved resource slices. The mobile terminals in Group 1 have mobile application flows, which occupied the mobile UE-slice. The number of content groups is *G*. We assume the total number of terminals in the scenarios is *S*. The total number of reserved resources in RBs is *N*. The reserved resource in RBs can be taken as one or more resource slices. In order to mitigate the collision, the silence probability is introduced as *Pq*. To evaluate the performance on latency and reliability of the proposed scheme, simulation scenarios are defined with a set of parameters {*G*, *S*, *N*, *Pq*, *M*, *R*}. We assume the value of TTI for one sub-frame is reduced to 100us, and the number of TTIs in a frame is 100. The length of retry window is *L* TTIs. The maximum number of retries is *R*. The maximum number of retries *R* can also be configurable due to the ultra-low latency limitation. Two performance metrics are utilized: the average delay *Ed* and the drop packet rate *Pd*.

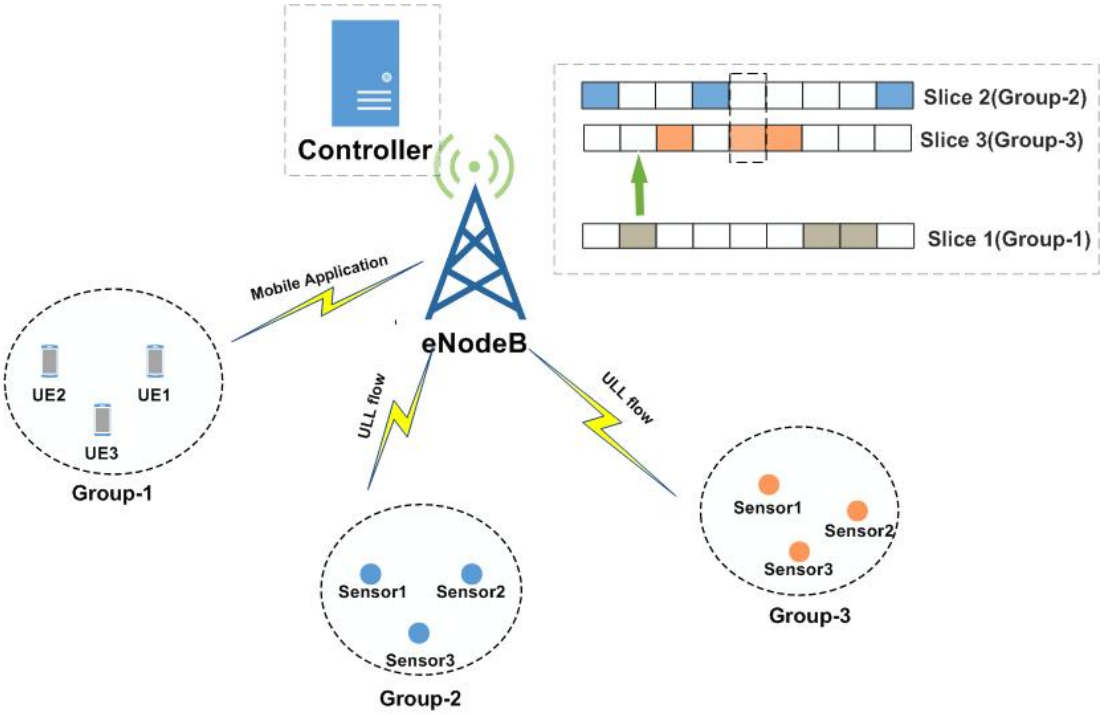

Fig. 5. Simulation scenario

4.2 Retransmission window

To meet the ultra-low latency in a few milliseconds, we analyze the impact of silence probability with the analytic model. We compare two pairs of configurations on the *Ed* and *Pd*, e.g. (*M*, *R*) is (100, 1), as shown in the **Fig. 6** and **Fig. 7**, and (20, 5), as shown in the **Fig. 8** and **Fig. 9**. In the **Fig. 6**, the number of terminals with emergency flow increases from 1 to 200 in one cell. We change the value of the silence probability Pq in terms of {0.1, 0.3, 0.5, 0.7 and

1.0}. The result on the average delay is shown in the **Fig. 6** with M=100 and R=1. The results show an average delay Ed increases in Pd for one content group. With the increase of Pq, the drop packet rate Pd also grows.

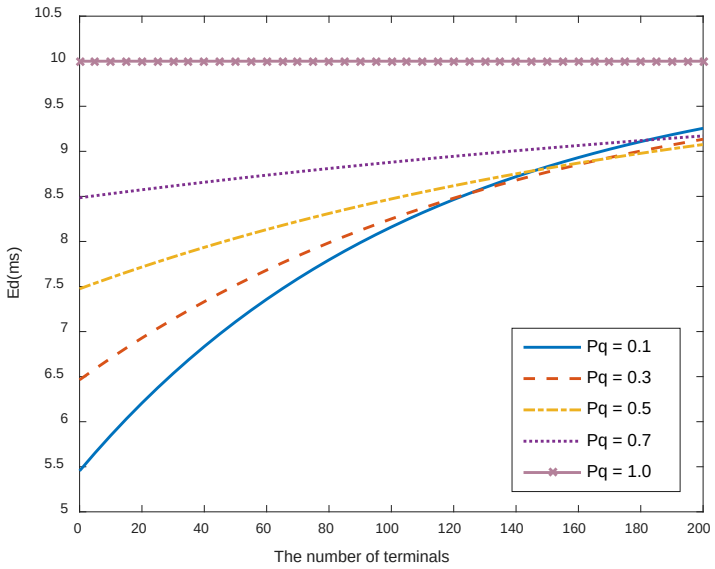

Fig. 6. The average delay Ed with R=1 and M=100

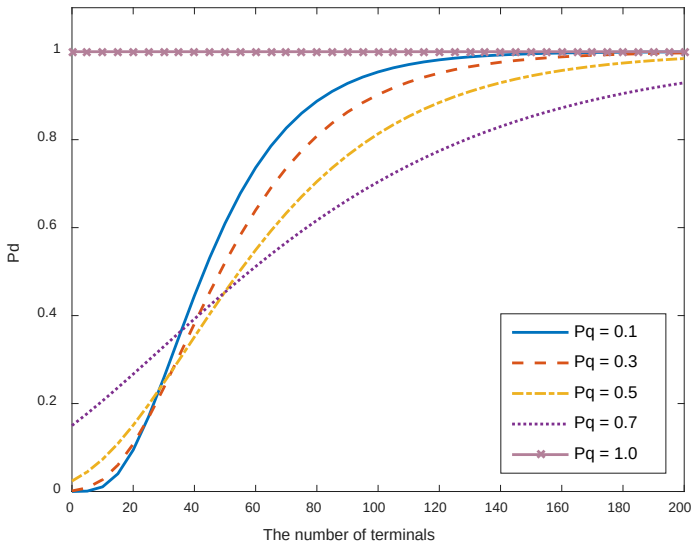

Fig. 7. The drop packet rate Pd with R=1 and M=100

The results shown in **Fig. 8** and **Fig. 9** with the M=20 and R=5. Compared with the results in the **Fig. 6** and **Fig. 7,** respectively, the performance of delay expectation and the drop packet rate are both improved. The new configuration of M=20 and R=5 is better, when the number of ultra-low latency flows is less than 30 in one cell. Therefore, the reduced window length introduces more retry opportunities, furthermore reduce the latency. Considering the requirement of ultra-low latency services in 5G, the optimal frame structure configuration for our scenario is R=5, M=20, and Pq=0.1.

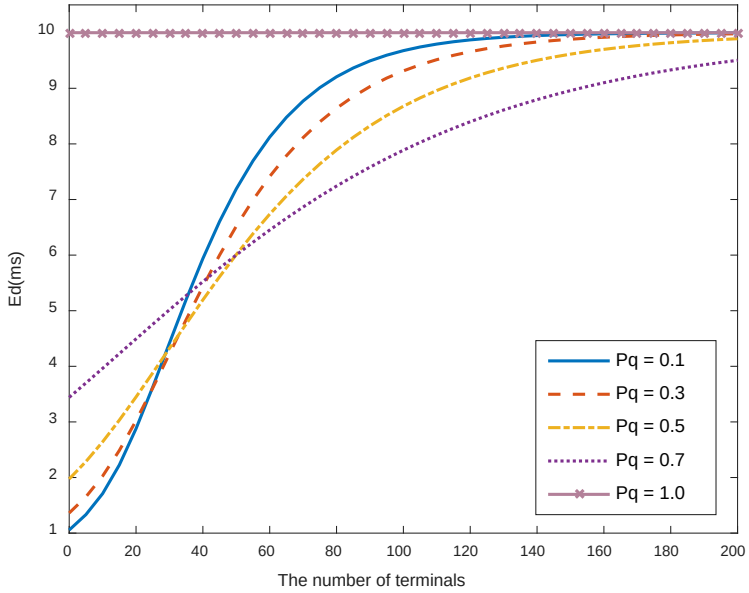

Fig. 8. The average delay E_d with R=5 and M=20

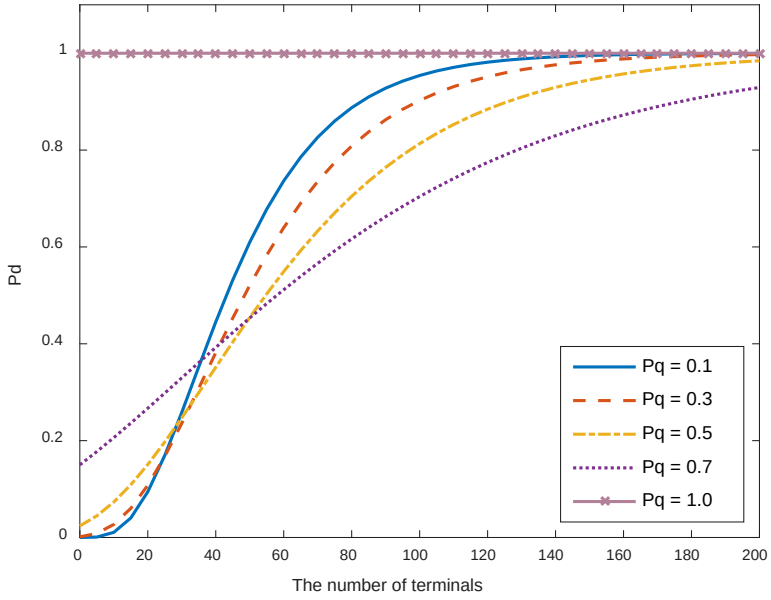

Fig. 9. The drop packet rate P_d with R=5 and M=20

4.2 Simulation vs. Analytic model

To validate the correctness of the proposed analytic model, Monte-Carlo simulations are performed and the numerical results are compared with the analytical results. The results are shown in the **Fig. 10** and **Fig. 11** with R=5, M=20, and Pq=0.1. We keep the analytic model in accordance with the simulation results. If the number of terminals with ultra-low latency flows is limited below 30, the average delay can be limited inside 3ms, and the drop packet rate can be limited under 1% with the configuration of M=20, R=5 and Pd=0.1.

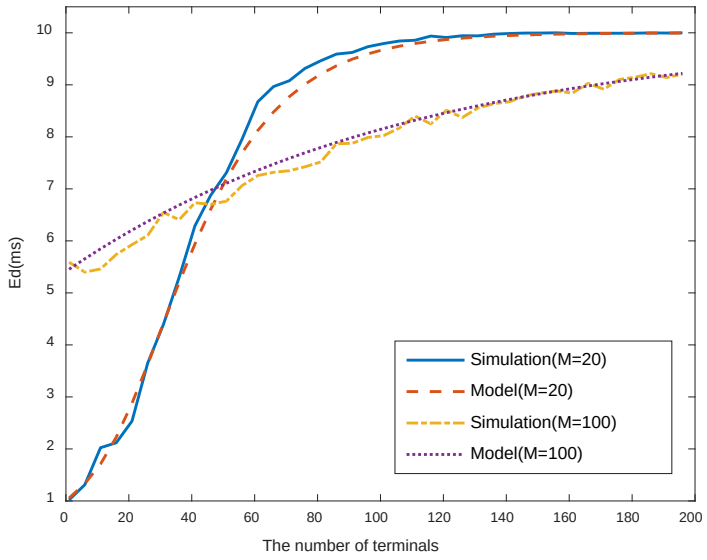

Fig. 10. The performance comparison on E_d

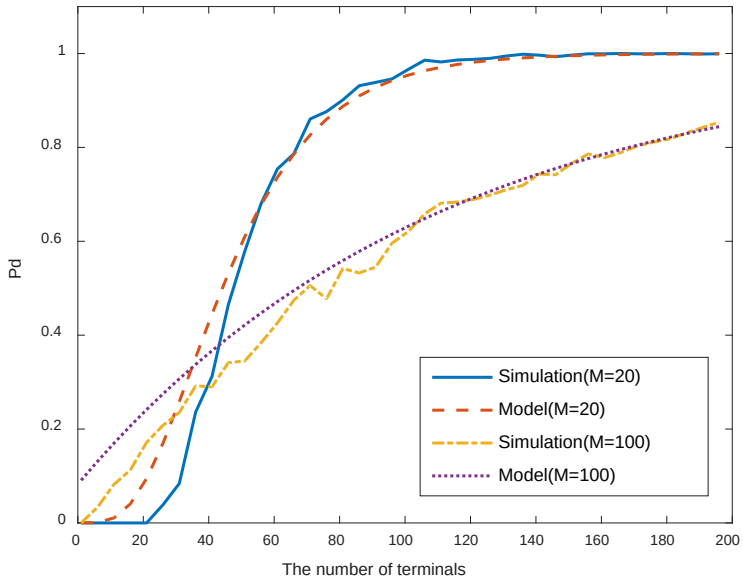

Fig. 11. The performance comparison on P_d

4.3 Extension to the multi-group

The result for one group reveals the feasibility to deploy 30 terminals with ultra-low latency flows in one cell. However, it is not enough in real cases. We extend this approach into multiple content groups in one cell with more RRs. Here, we consider two kinds of resource usage: Dedicated and Shared. In the **Fig. 12** and **Fig. 13**, the number of the groups, G changes from 1 to 3. In all six cases, the dedicated method provides a lower average delay when there are multiple groups, e.g. G=2 or 3. Similarly, the dedicated method outperforms the shared method and achieves a lower packet drop rate.

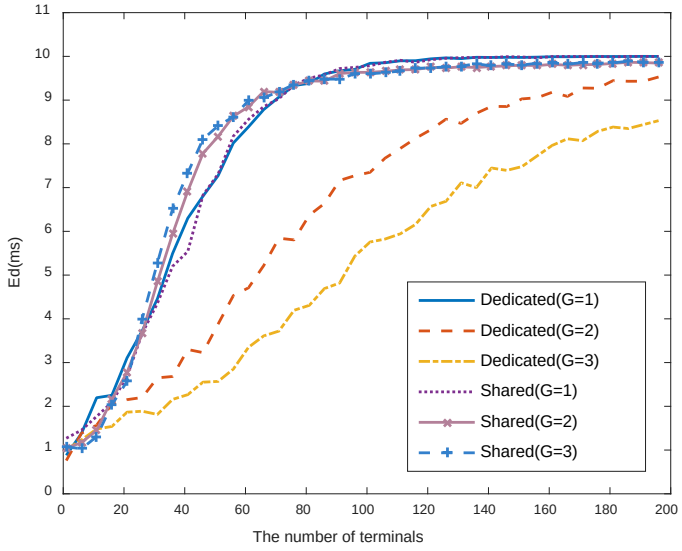

Fig. 12. E_d results for multiple groups

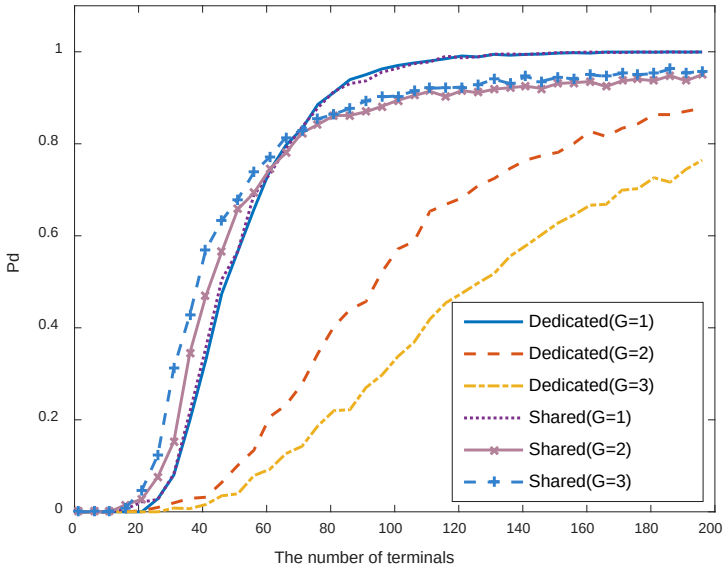

Fig. 13. P_d results for multiple groups

4.4 Persistent vs non-persistent resource reservation

Based on the simulation results above, the number of users supported on a single RB is almost thirty in the persistent resource reservation method. However, in this case, we consider the performance of the proposed method when the reserved resource in RBs is non-continuous. To compare the results in persistent reservation and non-persistent reservation, we set the parameters of simulation as follows. The length of TTI is 0.1ms, the size of retry the window L is 20, the limit of retransmission times R is 5, the silence probability is 0.1, and the number of terminals changes from 1 to 30. We run the simulation and compare the results in three cases of resource reservation. The first case is the original persistent resource reservation without

interval for RR region. The second is non-persistent, and the interval for each two neighbor RRs on time is 1 TTI. The third is non-persistent, and the interval for each two neighbored RRs is 2 TTIs. The results of the expected latency and the drop packet rate are collected and shown in **Fig. 14** and **Fig. 15**.

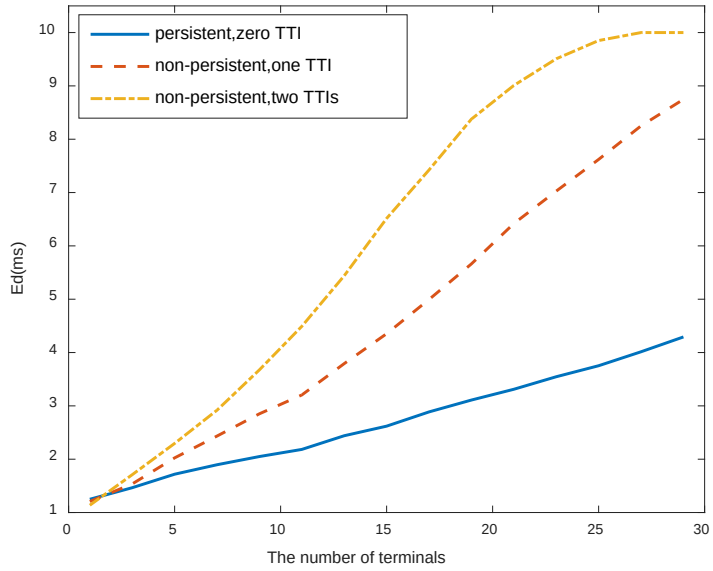

Fig. 14. E_d results for non-persistent resource reservation

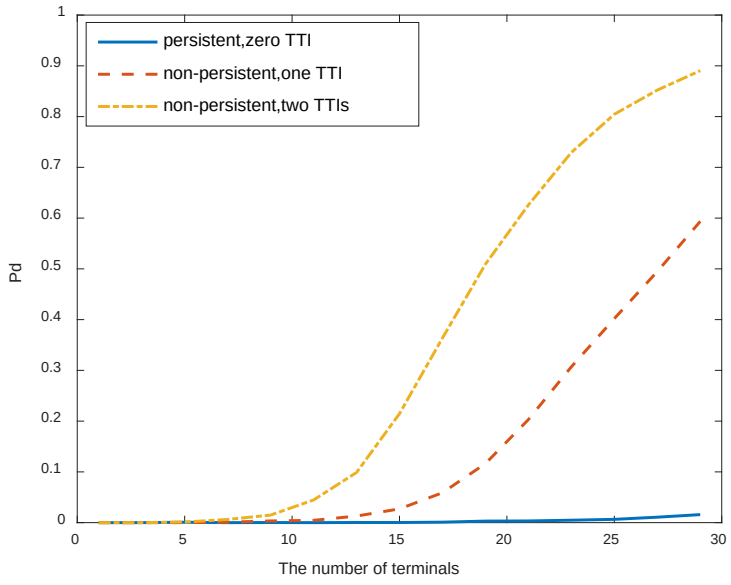

Fig. 15. P_d results for non-persistent resource reservation

The results in the **Fig. 14** and **Fig. 15** show that non-persistent resource reservation will decrease the performance on the latency and the packet drop rate. On the other hand, it shows that the RR region is really elastic and can adjust dynamically as one part of resource slice in the air-interface. Therefore, we discuss both the static and dynamic resource reservation.

4.5 Static vs dynamic resource reservation

We consider the performance of dynamic resource reservation for IoT terminals. The number of RRs in each retry window is changed dynamically with the estimated number of contention terminals. Because of the silence probability mechanism introduced, it equals to the sum of the number of collision terminals and the number of silent terminals. Furthermore, the number of TTIs with collisions T_{coll} in the retry window is counted by the base station. We compared the performance for latency and reliability with four resource reservation methods. The results are shown in **Fig. 16** and **Fig. 17**. The static method is the proposed resource reservation with one RB in each TTI. The dynamic resource reservation will adjust the number of RRs in the next retry window. The number of RRs can be limited, which depends on variable traffic load on the base station. Therefore, three sample cases are considered. Whether a resource has top limitation for each TTI: no-top limit, less than10 RBs, less than 5 RBs. The number of incoming terminals changes from 0 to 200. The other parameters: the window size in term of TTIs L=20, the number of retry times R=5, and the silence probability Pq=0.1.

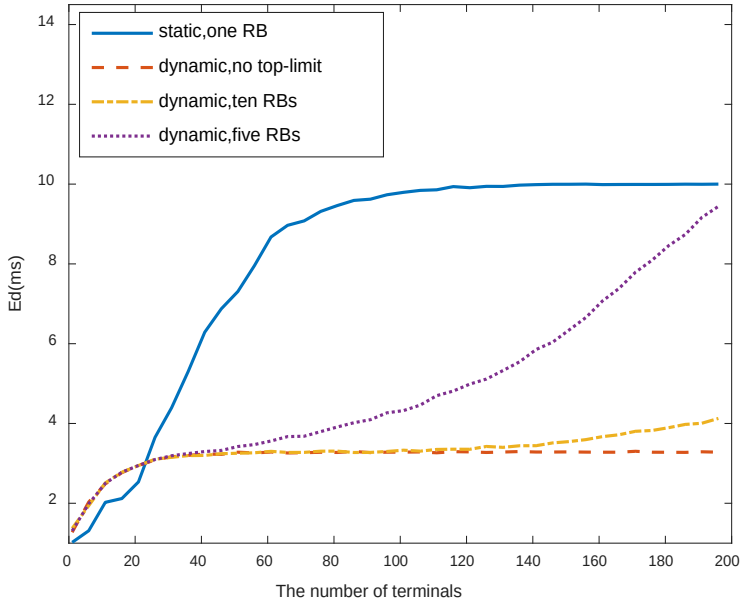

Fig. 16. E_d results for dynamic resource reservation

Based on these results, it is proved that dynamic resource reservation can improve the performance on the average delay and reliability. Based on the quantity, if the top-limit number of reserved RRs can reach five in each TTI, the average delay for 140 terminals will be nearly 3ms and the packet drop rate below one percent. If the top-limit number of reserved RRs is 10, the maximum number of supported terminals will be two hundred.

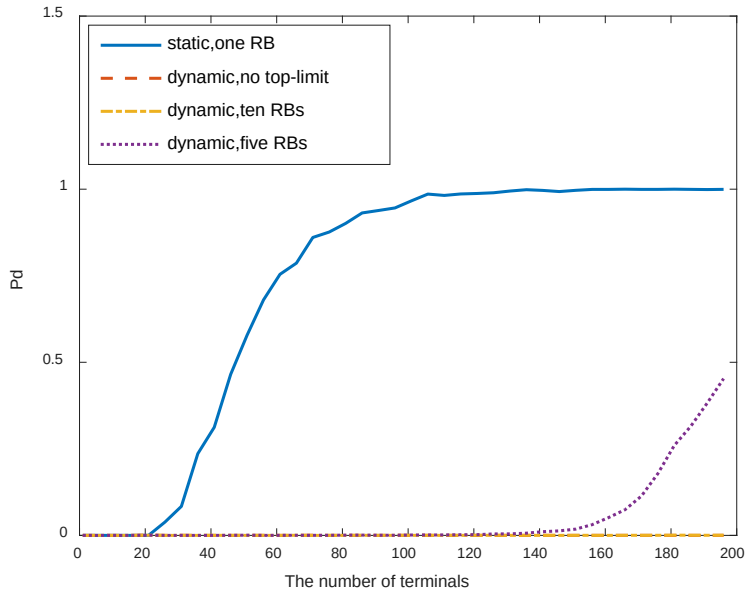

Fig. 17. P_d results for dynamic resource reservation

5. Conclusion

In this paper, a dynamic resource reservation schema based on air-interface slicing and a silence probability based collision mitigation algorithm are proposed for ultra-low latency IoT applications in 5G networks. A probability model is introduced to evaluate the performance of the proposed approach. The analytical and numerical results reveal that it can achieve an air-interface latency of a few milli-seconds and limit the packet drop rate in a few percentages. The dynamicity of resource configuration enables this approach to achieve an acceptable performance in the case of non-contiguous resource reservation and resource constrained scenarios. Further work includes the enhanced collision detection and estimation method on the reserved resources.

Acknowledgment

This work is supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China, Grant no. 20130185120021, by the Fundamental Research Funds for the Central Universities under grant no. ZYGX2014J060, and the ZTE Innovation Research Fund for Universities Program 2014 under grant no. CON1409180014.

References

[1] Popovski P, "Ultra-reliable communication in 5G wireless systems," in *Proc. of 1st IEEE International Conference on 5G for Ubiquitous Connectivity (5GU)*, pp.146-151, 2014. Article (CrossRef Link)

[2] Pablo Caballero Garces, Xavier Costa-Pérez, Konstantinos Samdanis, Albert Banchs, "RMSC: A Cell Slicing Controller for Virtualized Multi-Tenant Mobile Networks," in *Proc. of IEEE VTC Spring 2015*, pp. 1-6, 2015. Article (CrossRef Link)

[3] Mingju Li, Xiaoming She, Lan Chen, Otsuka, H., "A novel resource reservation scheme for fast and successful handover," in *Proc. of IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 556 – 560, 2009. Article (CrossRef Link)
[4] Adachi, K., Jingon Joung, Yuan Zhou, Sumei Sun, "A distributed resource reservation scheme for handover failure reduction," *IEEE Wireless Communications Letters*, vol. 4, Issue 5, pp. 537 – 540, 2015. Article (CrossRef Link)
[5] Shehada, M., Bo Fu, Thakolsri, S., Kellerer, W., "QoE-based resource reservation for unperceivable video quality fluctuation during handover in LTE," in *Proc. of IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 171 –177, 2013.
Article (CrossRef Link)
[6] Liu T C, Wang K, Ku C Y, et al, "QoS-aware resource management for multimedia traffic report systems over LTE-A," *Elsevier Computer Networks*, available online since Nov. 10, 2015.
Article (CrossRef Link)
[7] Corici, M.; Fiedler, J., Magedanz, T., Vingarzan, D., "Evolution of the resource reservation mechanisms for machine type communication over mobile broadband evolved packet core architecture," in *Proc. of IEEE GLOBECOM Workshops (GC Wkshps)*, pp.718 – 722, 2011.
Article (CrossRef Link)
[8] Albasheir S, Kadoch M, "Enhanced Control for Adaptive Resource Reservation of Guaranteed Services in LTE Networks," *IEEE Internet of Things Journal*, Vol. PP, Issue 99, pp.1-1, 2015.
Article (CrossRef Link)
[9] H. Egilmez, B. Gorkemli, A. Tekalp, and S. Civanlar, "Scalable video streaming over OpenFlow networks: An optimization framework for QoS routing," in *Proc. of ICIP 2011*, pp. 2241–2244, Sept. 2011. Article(CrossRef Link)
[10] Z. Zhu, S. Li, and X. Chen, "Design QoS-aware multi-path provisioning strategies for efficient cloud-assisted SVC video streaming to heterogeneous clients," *IEEE Trans. Multimedia*, vol.15, no.4, pp.758–768, Jun. 2013. Article(CrossRef Link)
[11] H. Egilmez, S. Civanlar, and A. Tekalp, "An optimization framework for QoS-enabled adaptive video streaming over OpenFlow networks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 710–715, Apr. 2013. Article(CrossRef Link)
[12] N. Xue, X. Chen, S. Li, L. Gong, D. Hu, and Z. Zhu, "Demonstration of OpenFlow-Controlled Network Orchestration for Adaptive SVC Video Manycast," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1617-1629, Sept. 2015. Article(CrossRef Link)
[13] Yaacoub, E., "On real-time smart meter reading using OFDMA based random access," in *Proc. of 17th IEEE Mediterranean Electro-technical Conference (MELECON)*, pp. 156 – 162, 2014. Article (CrossRef Link)
[14] Peng Xue, Hyunseok Ryu, Seong-Hoon Park, Sangwon Choi, "Collision-aware resource access in LTE-based device-to-device communication systems," in *Proc. of IEEE International Conference on Communication Workshop (ICCW)*, pp. 646-650, 2015. Article (CrossRef Link)
[15] Madueno G C, Stefanovic S, Popovski P, "Efficient LTE access with collision resolution for massive M2M communications," in *Proc. of IEEE Globecom Workshops (GC Wkshps)*, pp. 1433-1438, 2014. Article (CrossRef Link)
[16] Wu S, Chen Y, Chai K K, et al, "Analysis and performance evaluation of Dynamic Frame Slotted-ALOHA in wireless Machine-to-Machine networks with energy harvesting," in *Proc. of IEEE Globecom Workshops (GC Wkshps)*, pp.1081-1086, 2014. Article (CrossRef Link)
[17] Johansson N A, Wang Y P E, Eriksson E, et al, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. of IEEE International Conference on Communication Workshop(ICCW)*, pp.1184-1189, 2015. Article (CrossRef Link)
[18] Le Boudec J Y, Thiran P., "Network calculus: a theory of deterministic queuing systems for the Internet," *Springer Science & Business Media*, 2001.
[19] Feng B, Zhang C, Liu B, et al, "Sequentially ordered back-off:Towards implicit resource reservation for wireless LANs," in *Proc. of 2015 IEEE International Conference on Communications (ICC)*, pp. 3610-3615, 2015. Article (CrossRef Link)

[20] Sun, Guolin, Kefyalew, Dawit, Liu, Guisong,"Reservation based Resource Management for SDN-based UE Cloud," *accepted for publication in the KSII Transactions on Internet and Information Systems*.
[21] Ullah N, Ullah K, Islam S M R, et al, "Modeling MAC protocol based on frame slotted Aloha for low energy critical infrastructure sensor networks," *International Journal of Distributed Sensor Networks*, no. 13, pp. 1-11, 2015. Article (CrossRef Link)

Guolin Sun received his B.S., M.S. and Ph.D. degrees all in Comm. and Info. System from the University of Electronic Sci.&Tech. of China (UESTC), Chengdu, China, in 2000, 2003 and 2005 respectively. After Ph.D. graduation in 2005, Dr. Guolin has got eight years industrial work experiences on wireless research and development for LTE, Wi-Fi, Internet of Things (ZIGBEE and RFID, etc.), Cognitive radio, Location and navigation. Before he join the School of Computer Science and Engineering, University of Electronic Sci.&Tech. of China, as an Associate Professor on Aug. 2012, he worked in Huawei Technologies Sweden. Dr. Guolin Sun has filed over 30 patents, and published over 30 scientific conference and journal papers, acts as TPC member of conferences. Currently, he serves as a vice-chair of the 5G oriented cognitive radio SIG of the IEEE (Technical Committee on Cognitive Networks (TCCN) of the IEEE Communication Society. His general research interest is 5G/2020 oriented wireless network, including software defined networks, network function virtualization.

Guohui Wang received Bachelor of Computer Science from Hubei University, China in 2012. He is currently studying MSc in computer science at University of Electronic Science and Technology of China, due to finish in 2016. From 2013 to 2014, he worked as a Software Developer for Open Institute. His interest includes internet of things, SDN, and 5G.

Prince Clement Addo received his Bachelor in Information Technology from University Education, (College of Technology) Kumasi-Ghana, West Africa, in 2014. He is currently Studying MSc. Computer Science and Technology from the University of Electronic Science and Technology of China (UESTC). From 2005 to 2010, he worked as an Instructor and research assistance in Ghana Education Service and CCBTR UEW-K respectively. He is also a member of the Mobile Cloud-Net Research Team – UESTC. His interest include Mobile/Cloud Computing, Internet of Things, 5G and SDN.

Guisong Liu received his B.S. degree in Mechanics from the Xi'an Jiao Tong University, Xi'an, China, in 1995, M.S. degree in Automatics and Ph.D. degree in Computer Science both from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2000 and 2007 respectively. Now, he is an associated professor in the School of Computer Science and Engineering, UESTC. His research interests include cloud computing, big data, and computational intelligence.

Wei Jiang received his Ph.D degree from Beijing University of Posts and Telecommunications (BUPT) in 2008. Since Mar. 2008, he has been worked 4 years in Central Research Institute of Huawei Technologies, in the field of wireless communications and 3GPP standardization. In Sept. 2012, he joined the Institute of Digital Signal Processing, University of Duisburg-Essen, Germany, where he was a Postdoctoral researcher and worked for EU FP7 ABSOLUTE project and H2020 5G-PPP COHERENT project. Since Oct. 2015, he joined the Intelligent Networking Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, as a senior researcher and works for H2020 5G-PPP SELFNET project. Meanwhile, he also works for the Department of Electrical and Information Technology (EIT), Technische University (TU) Kaiserslautern, Germany, as a senior lecturer. He served as a vice Chair of IEEE TCCN special interest group (SIG) "Cognitive Radio in 5G". He is the author of more than 30 papers in top international journals and conference proceedings, and has 27 patent applications in wireless communications, most of which have already been authorized in China, Europe, United States or Japan. He wrote a chapter "From OFDM to FBMC: Principles and Comparisons" for the book "Signal Processing for 5G: Algorithms and Implementations" (Wiley, 2016). His present research interests are in digital signal processing, multi-antenna technology, cooperative communications, 5G, and machine learning