

Air-interface Slice Based Dynamic Resource Reservation for Ultra-low Latency IoT Transmissions

Guolin Sun, Guohui Wang, Guisong Liu
 School of Computer Science and Engineering
 University of Electronic Science and Technology of China
 Chengdu, China
 Email: guolin.sun@uestc.edu.cn

Abstract— The ultra-low latency transmission for emergency services needs an effective resources management scheme to deliver content in few milliseconds. The traditional solutions can't guarantee the ultra-low latency performance required by such traffic. In this paper, we propose an air-interface slice based dynamic resource reservation schema for a massive number of sensors with emergency flows in the context of the next generation cellular networks. The proposed schema allows ultra-low latency flows to be transported by guaranteed-rate radio link connection with a content name as identifier and it achieves air-interface latency in few milliseconds. The dynamic bit-map update, silence probability and window-based re-transmission are introduced based on the Frame Slotted Aloha, which can schedule the delay-sensitive flows immediately from one or many groups of connected terminals. Furthermore, a probability theory based analytic model is provided and evaluated with Monte-Carlo simulation results.

Index Terms— Ultra-low latency transmission; Network slice; Resource reservation

I. INTRODUCTION

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

- Remote control, data collection and environments sensing in the ultra-low latency applications, which often generates small packets in few coded bytes [1]. To support such massive small-packet transmissions, the complexity in radio resource allocation, scheduling and routing transfer is a big burden.
- These applications are extremely vulnerable to the end-to-end latency of data transmission. Providing ultra-low-latency services in few milliseconds consequently turns out to be the most crucial requirement.
- Traffic pattern of the emergency IoT flows is disruptive. Especially, the event happens occasionally for terminals in idle. The Radio Resource Control (RRC) re-establishment signaling overhead of terminals switching from the idle mode to a connected mode is a burden, in which eNodeB allocates

Resource Blocks (RBs) by UE's buffer status reporting.

To meet the diverse requirements from upcoming tactile applications, air-interface slicing for different quality of service class is proposed in 5G. Our basic motivation in this paper is to reserve some dedicated resources for ULL flows in order to mitigate overheads of control signaling. The data resources for the ULL flows will temporarily be borrowed from other slices, e.g. mobile-UE slice.

Till now, most of related works previously focused on the under-laying D2D communications in order to maximize the spectrum resource utilization to cellular network. The data collision problem in the distributed random access case has the critical impacts on the ultra-low latency services [2] [3] [4]. Recently, a reservation based collision aware resource access approach is proposed for the D2D communications [5]. The basic idea is to utilize the unused resource in data region for possible distributed coordination to avoid data collisions. The scheduling assignment region in a frame is used to reserve the resource blocks in the next frame. However, the scheduling assignment region is intermittent with the unused resource, which can mitigate collision but can't guarantee the ultra-low latency performance in few milliseconds. In [6], an initial analysis on ultra-low latency radio access problem for remote control is provided, which shows the reduced TTI (100us) without retransmission in the air interface is promising to achieve latency in one millisecond. In this paper, the proposed method is different with the conventional resource reservation. Firstly, this method is a persistent resource reservation, rather than a overlaying with the unused resource blocks. Secondly, the reserved resource is logic, which is dynamically updated for air-interface slices but temporarily scheduled out.

The organization of this paper is defined as follows. In the section II, a persistent and logic resource reservation approach with content grouping for RRC re-establishment is proposed to reduce the air interface latency. Furthermore, a probability theory based analytic model is given on the proposed approach. Performance evaluation is presented in Section III. Finally, we make conclusion in Section IV.

II. THE PROPOSED SCHEME AND ANALYSIS

A. The proposed approach

Resource reservation: To mitigate signaling burden when a terminal is switching from idle mode to a connected mode, a

persistent resource reservation method is adopted. Based on the traffic characteristics of IoT and D2D services, the packets are small. Furthermore, IoT can be taken as a content centric “database”. Therefore, we consider a content grouping based persistent and logical resource reservation in this paper. With a negotiation procedure, ultra-low latency application registers on the eNodeB or software defined controller. The negotiation message includes content name and a set of traffic parameters to identify the flow.

Air-interface slicing: As an example, the bandwidth resource of air-interface in 5G can be partitioned into two slices in a granularity of RBs: ULL-IoT slice and mobile-UE slice. The ULL-IoT slice is a logical definition in composed of virtual resource blocks. The virtual resource blocks will be mapped to physical RBs based on demand. The physical RBs for two different slices can be exchanged in a dynamic way on multiple slices. The RRs, illustrated in the Figure 1, is the reserved resource for ultra-low latency IoT flows scattered on the air-interface slice.

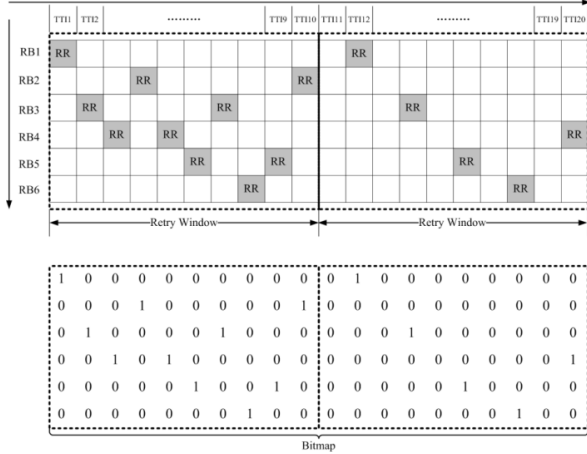


Fig. 1 The Bit-map design in a ULL-IoT slice

This resource indication in ultra-low latency IoT slice is sent to UEs in a broadcast sub-frame, which is configured as the first sub-frame in each retry window. This resource indication is based on a structure of bit-map, which includes the serial no. of slot and the RB index. Each terminal decodes the received bitmap information. Once an IoT terminal has the registered ultra-low latency data to send, this data will be transmitted on the reserved RBs and does not need to apply resources. The bit-map illustrated in the Figure 1 actually can be changed dynamically in each retry window, with collision detection and prediction of user population [7].

In this case, all of flows are aggregated in a content group and share a common bit-map based RR in a ULL-IoT slice, shown in the Figure 1. It is defined with a bitmap in a frame composed of some retry windows. The ULL flows occupy RRs based on a Frame Slotted Aloha. However, a collision may happen in this RR region. If collision didn't happen as shown in the Figure 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 Figure 2 to explain the specific behavior in signaling procedures.

The whole procedure can be summarized with four steps. Firstly, user can negotiate with the base station or controller whether it can utilize the reserved resource to send its flow. In the Figure 2, the initial number of reserved resource blocks on the ultra-low latency slice is assumed as two for each window. Second, the registered ultra-low latency flows can access the reserved RBs directly once it has data to transmit. However, if collision happened for more than one flows on some reserved RBs, they will retry in the next window. We assume collision happened between Flow 1 and Flow 2 in the Figure 2. Third, once the request got by the base station from the reserved RBs, the RBs on the ULL slice will be extended immediately by borrowing the resources from Non-ULL slice, as shown for F1 and F2 in the Figure 2. Fourth, the number of reserved RBs in each window is updated dynamically based on collision detection in the last window, which are always reserved for the new incoming ULL flows only. In the Figure 2, as a sample, the number of reserved RBs is assumed as 2, 2, 4 and 2 for each window.

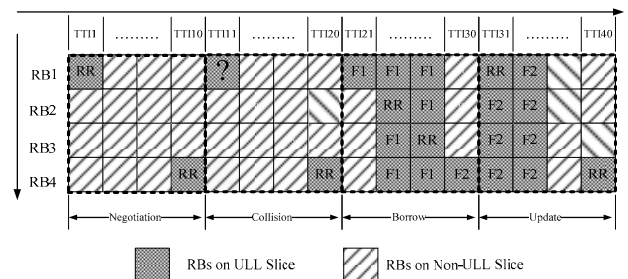


Fig. 2 The sliced resource view on the air-interface

Data contention: We assume that one frame is divided into M retry windows and collisions only happen in a content group. When the idle terminals have emergency data to transmit, they choose one of M sub-frames in random based on the indication in Master Information Block (MIB). The MIB embraces the reserved RB indication and the silence probability P_q . In order to mitigate collisions in the RR region, silence probability is inserted into the MIB message of every scheduling frame. This silence probability can be changed dynamically by a software defined controller based on traffic load. If a random value generated by this terminal is larger than a threshold, it keeps “silent” and retries this packet in the next window. Otherwise, it transmits it in one of M sub-frames. Once more than one packet is transmitted in a common RR region, collision happens. These terminals will re-transmit it again in the next retry window. The bitmap will be updated in the next retry window with collision detection and population estimation of users. When the number of retries reaches a maximum value, this packet will be dropped. This proposed collision resolution algorithm is illustrated in the Figure 3.

Extension to Multiple groups: The approach can be extended to multiple groups in the dedicated or shared ways. Regarding to the dedicated manner, 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, collision may happen among flows from two different content groups, but the resource utilization may be improved when the terminals are rare.

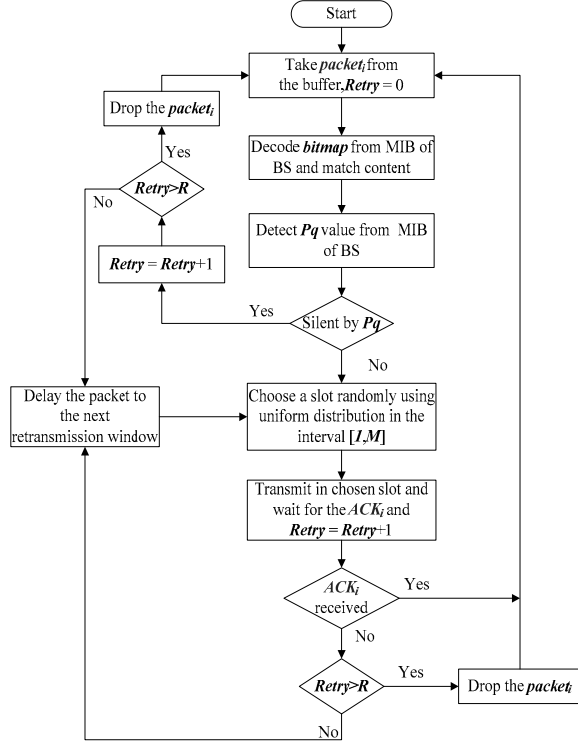


Fig. 3 The proposed contention resolution algorithm

B. Analytic Model

Definition 1: The success probability

Inside of a content group, if a terminal chooses one of M subframes to transmit, and the others with data to transmit choose the remaining $(M-1)$ subframes. Only in this case, the terminal can succeed in this transmission. 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

$$P_s = (1 - Pq) * \left(1 - \frac{1}{M}\right)^{N*(1-Pq)-1} \quad (1)$$

$$P_f = 1 - P_s \quad (2)$$

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 as follow [8].

$$P_d = P_f^R \quad (3)$$

Definition 3: Delay expectation

The delay expectation is defined as the average waiting time in buffer for a packet. The analytic model is illustrated in the Figure 4. For a terminal, there are R opportunities to succeed the transmission in the 1st to the R th packet. 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. The $P_s(i)$ can be calculated as

$$P_s(i) = P_f^{i-1} * P_s \quad (4)$$

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 E_d is:

$$E_d = \frac{1}{2} \sum_{i=1}^R P_s(i) * (i * T) + P_d * (R * T) \quad (5)$$

III. EXPERIMENTAL RESULTS AND ANALYSIS

In order to evaluate the performance on latency and reliability of the proposed approach, simulation scenarios are defined with a set of parameters $\{G, S, N, Pq, M, R\}$. The number of content groups is G , the number of terminals is S , the number of reserved resources is N , the silence probability is Pq . We assume the value of TTI for one sub-frame is reduced as 100us, and the number of TTIs in a frame is 100. Each retry window includes M 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 metrics are defined to evaluate the performance. One is the average delay, E_d . The other is the drop packet rate, P_d .

A. Re-try window

To control air-interface latency in few million-seconds, the impact of silence probability is evaluated with analytic model. We compare two pairs of configurations on the E_d and P_d , (20, 5), as shown in the Figure 4.

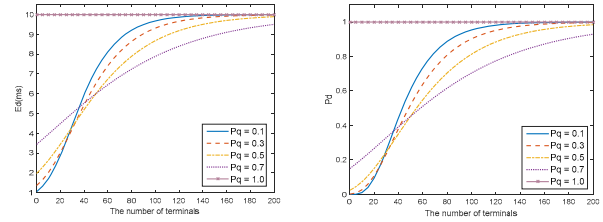


Fig. 4 E_d and P_d with $R=5$ and $M=20$

In the Figure 4, the number of terminals with emergency flow increases from 1 to 200 in one cell. We change the value of silence probability Pq in terms of $\{0.1, 0.3, 0.5, 0.7$ and $1.0\}$. With the configuration of $M=20$ and $R=5$, when the number of flows is less than 30 in a cell, it can achieve ultra-low latency performance in million seconds. Therefore, the reduced window length introduces more retry opportunities, furthermore reduces 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 $Pq=0.1$.

B. Simulation vs. Analytic model

To validate the correctness of the proposed analytic model, Monte-Carlo simulation is performed and compared with the analytic model based results. Results are shown in the Figure 5 with $R=5$, $M=20$, and $Pq=0.1$. The analytic model keeps in accord 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 $M=20$, $R=5$ and $Pd=0.1$.

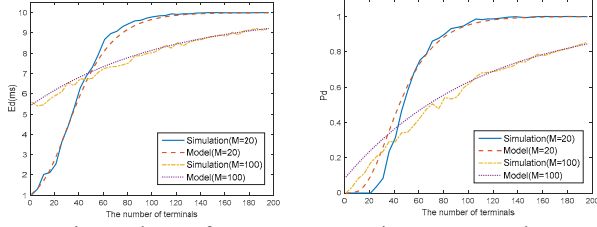


Fig. 5 The performance comparison on E_d and P_d

C. Extension to 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 Figure 6, the number of groups, G changes from 1 to 3. In these six cases, dedicated method provides a lower average delay when there are multiple groups, e.g. $G=2$ or 3. Similarly, dedicated method performs better than shared method and achieves a lower drop packet rate.

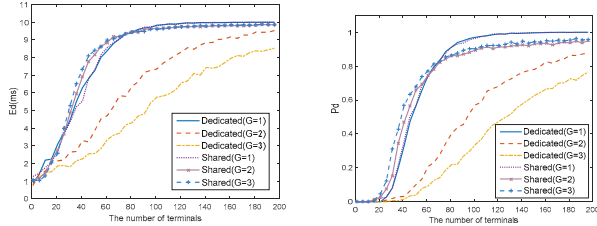


Fig. 6 E_d and P_d results for multiple groups

D. Persistent vs non-persistent resource reservation

Based on the simulation results above, the number of users supported on a single RB is almost 30 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, the length of TTI is 0.1ms, the size of retry window M is 20, the limit of retransmission times R is 5, the silence probability is 0.1, and the number of terminals is changing from 1 to 30.

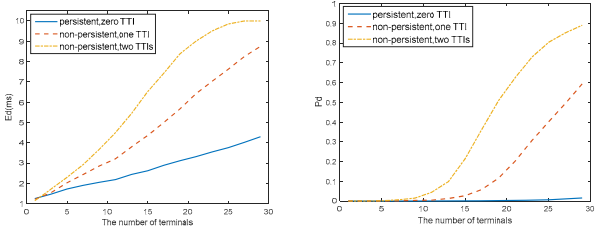


Fig. 7 E_d & P_d results for non-persistent resource reservation. We run the simulation and compare the results in three cases of resource reservation. One case is the original persistent resource reservation without interval for RR region. The second case is non-persistent, and the interval for each two neighbor RRs on time is one TTI. The third case is non-

persistent, and the interval for each two neighbored RRs is two TTIs. The results of the expectation latency and the drop packet rate are collected and shown in the Figure 7.

The results in the Figure 7 show that the non-persistent resource reservation approach will decrease the performance on latency and drop-packet rate. On the other hand, it shows us that the RR region is really elastic and can be adjusted dynamically as one part of resource slice in the air-interface. Therefore, we discuss static and dynamic reservation both.

IV. CONCLUSION

In this paper, an air-interface slice based resource reservation schema is proposed for ultra-low latency IoT transmissions in 5G. A persistent and logical resource reservation approach with a silence probability based collision mitigation algorithm is proposed. Then, an analytic model is provided for the proposed approach. Experiment results show that it reduces air-interface latency in few million seconds and the drop packet rate can be limited in few percentages.

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". *1st IEEE International Conference on 5G for Ubiquitous Connectivity*, 2014, pp.146-151.
- [2] Yaacoub, E., "On real-time smart meter reading using OFDMA based random access", *17th IEEE Mediterranean Electro-technical Conference*, 2014, pp. 156 – 162.
- [3] Madueno G C, Stefanovic S, Popovski P. "Efficient LTE access with collision resolution for massive M2M communications". *IEEE Globecom Workshops*, 2014, pp. 1433-1438.
- [4] 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", *IEEE Globecom Workshops*, 2014, pp.1081-1086.
- [5] Peng Xue, Hyunseok Ryu, Seong-Hoon Park, Sangwon Choi, "Collision-aware resource access in LTE-based device-to-device communication systems", *IEEE International Conference on Communication Workshop*, 2015, pp. 646-650.
- [6] Johansson N A, Wang Y P E, Eriksson E, et al. "Radio access for ultra-reliable and low-latency 5G communications", *IEEE International Conference on Communication Workshop*, 2015, pp.1184-1189.
- [7] Feng B, Zhang C, Liu B, et al, "Sequentially ordered back-off:Towards implicit resource reservation for wireless LANs," *2015 IEEE International Conference on Communications*, pp. 3610-3615.
- [8] 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*, 2015, no. 13, pp. 1-11.