The Design Methodology for MAC Strategies and Protocols Supporting Ultra-Low Delay Services in Next Generation IEEE 802.11 WLAN

Bo Li, Ghaleb Abdullah Abdulwahab Mohammed, Mao Yang(✉), and Zhongjiang Yan

School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
{libo.npu, yangmao, zhjyan}@nwpu.edu.cn,
abdullahl.ghaleb@gmail.com

Abstract. The next generation WiFi standard needs to consider how to better support ultra-low delay services. There are a lot of works proposed to improve the delay performance of traffic flows in WiFi networks. However, in order to face the high uncertainty of traffic arrival characteristics, it is necessary to explore new methodology to propose feasible Multiple Access Control (MAC) strategies and protocols supporting ultra-low delay services. This paper discusses the design methodology of ultra-low delay MAC strategies and protocols for next generation WiFi. Firstly, a general end-to-end transmission and processing model for an Information Transmission and Processing Network (ITPN) is proposed. The end-to-end delay of an ITPN is analyzed and the expression of the minimum end-to-end delay is obtained. Interestingly, based on the expression of the minimum end-to-end delay, we reveal three key factors that determine the end-to-end delay, namely, the number of processing blocks of the system, the size of information blocks processed and the total processing bandwidth of the system. Furthermore, some key technologies are proposed, which points out the feasible and attractive directions for the follow-up researches. Finally, a general ultra-low delay MAC framework based on the idea of "flexible reservation" is proposed. We believe that apart from IEEE 802.11 WLAN, the MAC framework proposed in this paper can be readily applied to various kinds of wireless networks.

Keywords: Wireless LAN · IEEE 802.11 · Ultra-low delay services · Medium access control · Reservation · Preemption

1 Introduction

WLAN technology regulated by IEEE 802.11 standard is developing continuously. As far as we know, the next generation IEEE 802.11 standard will support higher transmission rate and lower service delay requirements [1].

There are many researches on how to meet the low delay performance requirements for various time delay sensitive traffics in IEEE 802.11 WLAN. In general, the related works can be divided into two categories (see "related work" for details): works to

B. Li et al. (Eds.): IoTaaS 2020, LNICST 346, pp. 80–97, 2021.
https://doi.org/10.1007/978-3-030-67514-1_7

reduce service delay [2–23] and works to eliminate service delay [24–32]. In order to reduce service delay, authors try to improve the performance of service delay by optimizing scheduling strategy [6–13], optimizing queuing strategy [2–5], and distinguishing access priority [14–19]. We believe that these methods are not suitable for supporting the performance requirements of ultra-low delay services. On the other hand, resource reservation and resource preemption are the key methods to eliminate service delay. Because these methods try to eliminate service delay as much as possible, it is believed that they are more suitable for supporting ultra-low delay traffics.

However, considering the uncertainty of traffic arrival characteristics, the medium access strategies based on resource reservation cannot accurately predict the arrival characteristics of a traffic flow, and thus cannot achieve efficient resource reservation. For the medium access strategies based on resource preemption, the preempted traffic flows may be starved by traffic flows with higher priority. Therefore, in view of the uncertainty of traffic arrival characteristics, how to support ultra-low delay service efficiently is one of the key problems to be solved in designing and implementing the next generation WiFi medium access technologies.

This paper discusses the design methodology of ultra-low delay medium access control strategies and protocols for next generation WiFi. The main contributions include:

- A general end-to-end transmission and processing model for an ITPN is proposed. Based on this model, the end-to-end delay of the system is analyzed and the expression of the minimum end-to-end delay is given. Three key factors that determine the end-to-end delay are revealed, that is, the number of processing blocks, the size of information blocks processed and the total processing bandwidth of the system.
- Some key technologies and the core ideas to realize ultra-low delay services are proposed, which points out the feasible directions for the follow-up researches.
- A general ultra-low delay MAC framework based on flexible reservation is outlined. The MAC framework can be applied to various kinds of wireless communication networks including IEEE 802.11 WLAN.

The paper is organized as follows: in Sect. 2, a brief overview of the existing works aimed at improving the service delay performance in IEEE 802.11 networks is given. In Sect. 3, the end-to-end delay in an ITPN system is generalized, modeled and analyzed. In Sect. 4, basic ideas for the key technologies to supporting ultra-low delay are described. In Sect. 5, a MAC framework for supporting ultra-low delay services is proposed. In Sect. 6, some conclusions are given.

2 Related Work

2.1 Latency Reduce Scheme

A. Optimizing Queuing Strategy

Packet queue is an indispensable module in the media access control (MAC) layer of WiFi devices. Packets from the upper layer to the MAC layer often need to be assigned

to one data queue. According to Enhanced Distribution Channel Access (EDCA) mechanism specified in IEEE 802.11e, WiFi device always possesses several queues, four or six typically. Each queue corresponds to one access category (AC). It is obvious that packets need to experience queuing delay before being transmitted. Therefore, some studies focus on queuing strategy optimizing for low latency.

Saheb, and et al. [2] propose an enhanced hybrid coordination function (HCF) controlled channel access (EHCCA) scheme. In EHCCA, the packets are queued in MAC based on the traffic deadline. In this case, more urgent packets can be scheduled due to higher priority. Pei, and et al. [3] allows the late arriving packet with low latency requirement to bypass the data queuing. Li, and et al. [4] introduce a new parameter that is related to packet importance index, packet latency, and channel state, where the packet importance index indicated in the IP header can be derived based on gradient of the video QoE function. After that, naturally, the enqueue operation is directly affected by the importance index. Prabhu, and et al. [5] face to the polling based IEEE 802.11 network. The authors analyze that with the increase of STA number, the latency performance is getting worse. Then, two-level priority queues are proposed.

B. Optimizing Scheduling Strategy

Scheduling is quite important for wireless networking and communications. Ahn, and et al. [6] propose a MAC scheme based on Orthogonal Frequency Division Multiple Access (OFDMA) introduced by IEEE 802.11ax. The Access Point (AP) periodically sends trigger frame (TF) to require the associated low latency STAs to send uplink data through OFDMA without access collision. Qian, and et al. [7] introduce an aggregation sliding window for aggregate MAC protocol data unit (A-MPDU) even if some MPDUs aggregated in the A-MPDU exceed the bitmap range. To determine the aggregated MPDU number in the retransmission A-MPDU, the proposed scheme takes the delay requirements into consideration since large A-MPDU duration may lead to large latency. Some other studies also focus on the A-MPDU optimization such as aggregation size, per-MPDU size, and retransmission time [8–11].

Avdotin, and et al. [12, 13] try to enhance the uplink OFDMA random access (UORA) scheme specified in IEEE 802.11ax. The authors propose two schemes. In the first scheme, AP allocates several resource units (RUs) for low latency requirements only. When collision occurs in these RUs, AP polls the low latency STAs one by one in the next several successive TFs. In the second scheme, the low latency STAs are divided into several groups. The STAs in the same group are allocated in one RU. If collision occurs in one RU, AP polls the STAs in the corresponding group in the next few TFs.

C. Distinguishing Access Priority

IEEE 802.11 adopts carrier-sense multiple access with collision avoidance (CSMA/CA). It means contention based channel access is deployed before transmission. Channel access delay cannot be ignored especially when node number is large.

Kim, and et al. [14] modify the channel access rules by introducing "Round". Round is a time duration during which each STA can at most send one frame. They validate that the proposed scheme improves the tail latency for IEEE 802.11 network. Nguyen, and et al. [15] analyze the requirement differences between delay sensitive traffic and throughput sensitive traffic. With the change from delay sensitive to throughput sensitive, both the contention window (CW_{min}) and transmission opportunity (TXOP) limit proportionally increase. Tian, and et al. [16] modify the backoff rule that not only the backoff value is doubled due to transmission failure, but also the backoff value is doubled when the channel is busy. Lin, and et al. [17] model the system performance with all the associated STA's channel access configuration and parameters by using machine learning. Then, based on the trained model, AP optimizes the channel access parameters of each STA and announces these parameters through beacon frame to obtain the delay requirements. Moreover, some studies focus on the contention window size adjusting algorithm [18, 19].

D. Others

Besides the schemes discussed above, some studies try to improve the latency from other perspectives. Some related work introduces redundancy resources to enhance the transmission reliability, thereby improving the latency performance [20, 21]. Moreover, some researchers focus on the static configuration such as operation channel, transmission power, location, and load balance approaches [22, 23].

2.2 Latency Elimination Schemes

A. Channel Reservation

Channel reservation allows the prior transmission piggybacks the transmission time and duration of the subsequent transmissions. Other nodes who successfully receive the reservation information will keep silent during the reserved period. Choi, and et al. [24] propose a reservation scheme named EBA. One node broadcasts the backoff value of the next channel access process. Other nodes obtaining this information avoid selecting the same backoff end time.

Our previous studies extend the channel reservation conception into multi-step channel reservation [25, 26]. We analyze that one-step reservation is not reliable due to channel loss and collisions, resulting in the failed transmission of reservation information. We propose that one packet piggybacks the reservation information of several subsequent packets, named multi-step reservation. This scheme significantly enhances the transmission reliability of reservation information since the reservation information of each packet is transmitted by several times.

Some studies focus on the reservation schemes in the wireless ad hoc network and mesh network. Singh, and et al. [27] pay attention to the periodical traffic. The nodes periodically access the channel, and let other nodes know their periods to avoid the collisions. Sheu and T. Sheu [28, 29] propose two-period reservation MAC protocol.

Nodes in the contention based period transmit data or broadcast the reservation information, while in the contention free period they directly transmit data in the reserved time slot without contention. Moreover, other studies introduce the reservation schemes in synchronized network [30, 31].

B. **Resource Preemption**

Resource preemption allows the intended low latency transmission to pause, stop, or preempt the ongoing transmission. Bankov, and et al. [32] propose an out-of-band announcement based preemption scheme. The STA who has intended low latency transmission firstly sends an announcement frame in a dedicated channel. Then, the sender of the ongoing transmission stops transmission immediately, and meanwhile, other STAs suspend their backoff process. After that, the STA who has intended low latency transmission can send its frame immediately.

3 Modeling and Analysis of End-to-End Delay in Information Transmission and Processing Networks

IEEE 802.11 WLAN is a kind of information transmission and processing network. In order to make the methodology proposed in this paper more general, we summarize a general end-to-end delay model for an ITPN, and analyze the general expression of the end-to-end delay based on it. Then, given the total processing bandwidth of the system, the expression of the minimum end-to-end delay is obtained.

3.1 General System Model of an ITPN

Figure 1 shows a simple ITPN, that is, there are only two nodes (sending node and receiving node) in the system. Furthermore, we can decompose each node into several protocol layers. In other words, information originates from the protocol layer N of the sender and goes through the protocol layers $N, N-1, \ldots, 1$ of the sender in turn. Then, the information processed by the sender is transmitted into the protocol layer 1 of the receiver, and from the protocol layer 1 of the receiver, it goes through the protocol layers $1, \ldots, N-1, N$ of the receiver in turn. Suppose that the processing delay of the message to be processed at the protocol layer i of the sender is $T_{ds}(i)\ (i = 1, 2, \ldots, N)$ and that of the protocol layer j of the receiver is $T_{dr}(j)\ (j = 1, 2, \ldots, N)$. It can be concluded that the end-to-end delay experienced by the information in this simple ITPN is as follows:

$$T_D = \sum_{i=1}^{N} T_{ds}(i) + \sum_{j=1}^{N} T_{dr}(j) \tag{1}$$

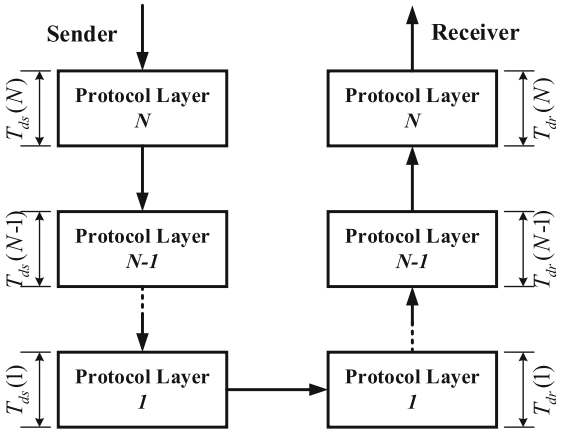


Fig. 1. Two communication terminals with layered protocol stacks

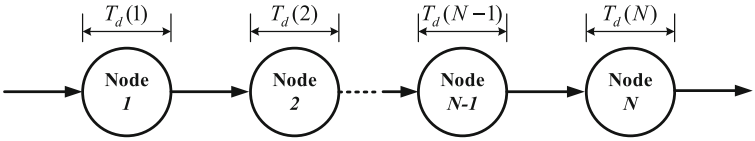


Fig. 2. A multi-hop network composed of several nodes

Figure 2 shows another example of an ITPN, that is, the system contains N adjacent communication nodes, and the information is processed and transmitted hop-by-hop in the way of storage and forwarding. In other words, the information starts from node 1, goes through the processing of the nodes $1, \ldots, N-1, N$ in turn, and finally is output at the node N. If the processing delay of the processed information in the node i is $T_d(i)$ $(i = 1, 2, \ldots, N)$, we can obtain that the end-to-end delay of the information in the system is:

$$T_D = \sum_{i=1}^{N} T_d(i) \tag{2}$$

Above, we only give two examples for ITPN systems (there are still many more). If we generally perceive both "protocol layer" in Fig. 1 and "node" in Fig. 2 as "processing block", we will get a more general end-to-end processing and transmission model for an ITPN as shown in Fig. 3. It can be seen from the model that most of ITPNs can be generally considered as a network system which processes and transmits input information flow from one block to another. Therefore, the delay experienced by

the input information in an ITPN can be considered as the superposition of processing delays caused by each processing block

$$T_D = \sum_{i=1}^{N} T_d(i) \tag{3}$$

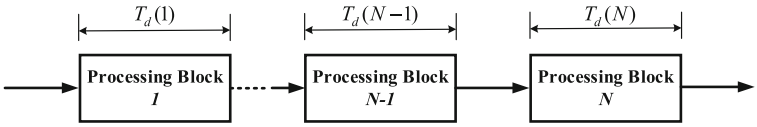


Fig. 3. A general system model for an ITPN

3.2 Modeling of End-to-End Delay

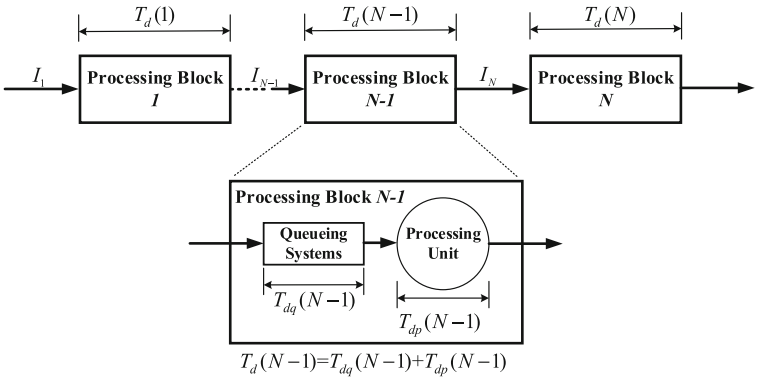


Fig. 4. The decomposition of the end-to-end delay in an ITPN

As shown in Fig. 4, the delay $T_d(i)\ (i = 1, 2, \ldots, N)$ in each processing block can be further decomposed into queuing delay $T_{dq}(i)\ (i = 1, 2, \ldots, N)$ and processing delay $T_{dp}(i)\ (i = 1, 2, \ldots, N)$. Therefore, the end-to-end delay in an ITPN can be expressed as:

$$T_D = \sum_{i=1}^{N} [T_{dq}(i) + T_{dp}(i)] \tag{4}$$

It is worth noting that the queuing delay $T_{dq}(i)\ (i = 1, 2, \ldots, N)$ is the sum of all the waiting times of the processed information waiting for further processing in the current processing block.

Define the processing bandwidth $b_p(i)\,(i = 1, 2, \ldots, N)$ of the current processing block as:

$$b_p(i) \triangleq \frac{I_i}{T_{dp}(i)} \tag{5}$$

where, $I_i\,(i = 1, 2, \ldots, N)$ represents the amount of information to be processed by the current processing block (unit: bits). Combined with (4) and (5), we have

$$T_D = \sum_{i=1}^{N} \left[T_{dq}(i) + \frac{I_i}{b_p(i)} \right] \tag{6}$$

3.3 Minimization of End-to-End Delay

In this paper, we consider how to achieve ultra-low end-to-end delay. If we can take some measures (such as the related technologies described in Sect. 4) to eliminate queuing delays, i.e. let $T_{dq}(i) = 0\,(i = 1, 2, \ldots, N)$, the end-to-end delay in an ITPN can be simplified as follows:

$$T_D = \sum_{i=1}^{N} \left[\frac{I_i}{b_p(i)} \right] \tag{7}$$

Next, on the basis of Eq. (7), we consider how to minimize the end-to-end delay T_D and what is the value of the minimum end-to-end delay $T_{D,\min}$. It is assumed that the amount of information input into an ITPN remains unchanged after being processed from one processing block to another, i.e. $I = I_i\,(i = 1, 2, \ldots, N)$. And given the total processing bandwidth of the system to be $B = \sum_{i=1}^{N} b_p(i)$, the optimization analysis of Eq. (7) shows that when the total processing bandwidth of the system is equally allocated to each processing block, the corresponding end-to-end delay reaches the minimum value. That is,

$$T_{D,\min} = \frac{I \cdot N^2}{B} \tag{8}$$

In order to facilitate the readers' understanding, we list some minimum end-to-end delays $T_{D,\min}$ corresponding to different parameters I, N, B in Table 1. Do you think which performance level can be more practically achieved by the current wireless networks?

Table 1. Examples of end-to-end delays achieved

N: Number of processing blocks	I: Size of an information block (bits)	B: Total processing bandwidth (Mbps)	Minimum total end to end delays ($T_{D,\min}$:ms)
100	1,000	10,000	1.0
100	100	1,000	1.0
10	100	10	1.0
10	100	10,000	0.001

4 Basic Ideas of the Key Technologies Supporting Ultra-Low Delay Services

In the previous section, we analyzed and modeled the end-to-end delay in an ITPN, and analyzed the minimum end-to-end delay that can be achieved for an ultra-low delay traffic flows. In the analysis, we assume that the queuing delay of an ultra-low delay traffic flow in each processing block is zero. However, in order to achieve zero queuing delay and minimize processing delay, a variety of key technologies are indispensable. In this section, we will focus on the core objectives of eliminating queuing delay and/or reducing processing delay, and introduce the basic ideas for several key technologies, which is summarized in Table 2.

Table 2. Basic ideas of key technologies

Key technologies	Basic ideas
Traffic identification	When it enters into an ITPN, the corresponding traffic flow will be identified and processed
Simplification of processing tasks	For ultra-low delay traffic flows, fewer processing blocks will be provided as much as possible
Instant processing	The information arriving successively is organized into the information unit with the smallest granularity as possible. The corresponding processing block can start its processing instantly
Resource preemption	When a traffic flow with ultra-low delay requirement arrives at the current processing block, if there is no idle resources for it, it can preempt the processing resources occupied by other delay insensitive traffic flows
Resource reservation	Based on the prediction of the arrival characteristics of an ultra-low delay traffic flow in the future period, corresponding processing resources are reserved for it in advance
Traffic prediction	The possible amount of information to be processed for an arrival traffic flow in a certain future period is estimated in advance

(*continued*)

Table 2. (*continued*)

Key technologies	Basic ideas
Flexible reservation	Through the efficient combination of resource reservation, resource preemption and resource release, a flexible "exchange" mechanism of occupied resources is introduced (with resource release and resource preemption as the means of exchange) between traffic flows with different performance requirements, and overcomes the intrinsic defects which cannot be easily solved by using only resource reservation or resource preemption
Traffic flexibilization	Reducing the sensitivity of the loss of the utility gain for the recipient of a traffic flow to the loss of partial information in the traffic flow
Grouping based scheduling for traffic flows	The problem to be solved is how to effectively group traffic flows and map specific traffic groups to specific processing bandwidth. The core idea is to organize some class B traffic flows with greater flexibility and some class A traffic flows, which is difficult to accurately predict their arrival characteristics, into some common group to be processed further within a shared processing bandwidth
Collaborative reservation	The basic idea is that when an earlier processing block clearly knows the arrival characteristics of a traffic flow, it informs the subsequent processing blocks to reserve the required amount of processing resources in advance through the signaling system between processing blocks
Adaptive allocation of processing bandwidth	Based on the amount of information to be processed, the corresponding processing bandwidth is allocated adaptively
Optimal configuration of processing bandwidth	Under the given constraints, the end-to-end delay can be minimized by optimizing the allocation of processing resources among all the processing blocks in an ITPN

4.1 Traffic Identification

In an ITPN, it often carries several traffic flows with different performance requirements (i.e. the traffic flows that need to be transmitted and processed by the ITPN). The system needs to take some technical measures for traffic flows with ultra-low delay performance requirements (see the descriptions below). Therefore, it is necessary to identify the traffic flow as soon as it enters the ITPN, so as to configure the processing strategies used by each processing block. In short, the central task of the traffic identification is to identify the traffic flow when it just enters the ITPN, and instruct other processing blocks to properly configure their processing strategies.

4.2 Simplification of Processing Tasks

In an ITPN, a variety of traffic flows need to be processed by several processing blocks, and each processing block brings about a certain delay (including queuing delay and

processing delay). It can be seen from Eq. (8) that the minimum end-to-end delay will increase significantly with the number of processing blocks. In order to reduce the delay, an obvious idea is to reduce the number of processing blocks as much as possible for the traffic with ultra-low delay requirements (that is, the fewer processing blocks, the better!). For example, for some ultra-low delay traffic flows, one can consider using a simplified network protocol stack to reduce the number of protocol stacks that must be passed through. As for another example, in a multi-hop ad hoc network, the end-to-end routing with as fewer hops as possible is established for ultra-low delay traffic, so as to reduce the number of "store and forward" links that the traffic must go through in the transmission procedure. In short, the core idea is to provide as fewer processing blocks as possible for ultra-low delay traffic flows, so as to achieve the purpose of reducing end-to-end delay.

4.3 Instant Processing

The information of a traffic flow to be processed in an ITPN is often divided into information units, and then corresponding processing is carried out for each information unit. If the granularity of the information unit is large, it will lead to a large waiting delay before the start of the processing. It can be seen from Eq. (8) that the minimum end-to-end delay will increase linearly with the increase of the size of information units processed by a processing block. In other words, if the size of the information unit can be reduced as much as possible, the end-to-end delay will be reduced accordingly. The basic idea of instant processing is to organize the information flow that reaches the ITPN system successively into information units with the smallest granularity as possible, and start the processing instantly without waiting for a longer period of time. The idea of instant processing can be applied in various scenarios, such as instant framing technology, instant coding technology, and instant medium access control technology and so on.

4.4 Resource Preemption

When an ultra-low delay traffic flow reaches the current processing block, if the processing unit in the processing block has no available processing resources (that is, all its processing bandwidth is occupied by other traffic flows), the ultra-low delay traffic flow can only wait in the queue. The basic idea of resource preemption is that when the ultra-low delay traffic flow arrives, if there is no idle processing resources, it can preempt the processing resources occupied by the other delay insensitive traffic flows. The preempted traffic flows will be temporarily arranged in the queue of processing block, waiting for subsequent resources to be available so as to continue their suspended services. The essence for resource preemption is to reduce the processing delay of ultra-low latency traffic flows at the cost of extending the delay of insensitive traffic flows.

4.5 Resource Reservation

Resource preemption technology has a certain cost, which not only prolongs the delay of the preempted traffic flows, but also causes low priority services to be possibly "starved" by ultra-low delay traffic flows without exerting some proper measures. Moreover, it will cost a certain amount of time to process these interrupts. Resource reservation is a technology that can react faster than resource preemption. Its basic idea is to reserve processing resources in advance according to the prediction of the arrival characteristics of an ultra-low delay traffic flow in the future period (the reserved resources will not be occupied by other traffic flows). In this way, when the ultra-low delay traffic flow arrives, the reserved resources can be used instantly without affecting the processing resources occupied by other services.

4.6 Traffic Prediction

The premise for resource reservation is to accurately predict the ultra-low delay traffic, that is, to estimate the possible amount of traffic arrivals for the corresponding ultra-low delay traffic flow in a certain future period, so as to reserve proper amount of resources accordingly. Generally speaking, the arrival characteristics of most traffic flows are uncertain to some different extent, which makes it an extremely challenging task to achieve accurate traffic prediction. Hence, this implies that it is also very challenging to realize accurate reservation of resources!

4.7 Flexible Reservation

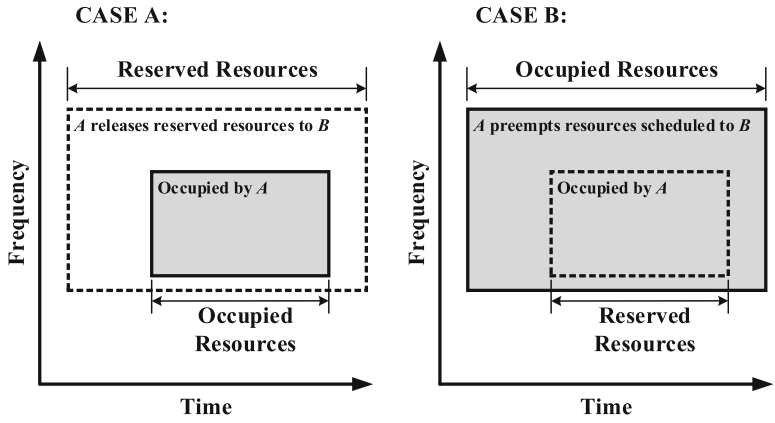


Fig. 5. Flexible reservation mechanism

The basic idea of flexible reservation is to reasonably combine resource reservation, resource preemption and resource release, and introduce an efficient "exchange" mechanism (with resource release and resource preemption as the means for making exchange) for occupied resources between traffic flows with different performance

requirements, and then overcome the intrinsic defects that both resource reservation and resource preemption cannot overcome individually. Figure 5 shows the core idea of flexible reservation. For the sake of convenience, we classified traffic in an ITPN into time delay sensitive traffic (class A traffic) and delay insensitive traffic (class B traffic). For a class A traffic flow, it normally relies on resource reservation to enjoy the processing of the system. If the reserved resources are not actually used by it (due to inaccurate traffic prediction), these resources can be released and used by other class B traffic flows (as shown in case A in the figure). On the other hand, for a class A traffic flow, it actually needs to use processing resources that have not been reserved in advance (again due to inaccurate traffic prediction), and in this case it can preempt the resources that are being occupied by some class B traffic flows (as shown in case B in the figure). It can be seen that the flexible reservation mechanism puts much lower requirements on the accuracy for making traffic prediction, which makes it to be more feasible and practical than the above mentioned resource reservation mechanism.

4.8 Traffic Flexibilization

In an ITPN system, the recipient of a traffic flow will get a certain utility gain after receiving a certain amount of information. If not all the information to be received reaches the recipient before the expected time instance (that is, only a part of the information arrives, and the part of information that fails to arrive on time will lose its effectiveness), the utility gain of the recipient will be lost to some extent. The flexibility of a traffic flow means that before the expected time instance, although some information has not reached to the recipient, the loss of the utility gain that the recipient obtained is not large. In other words, the loss of the utility gain at the recipient side is not sensitive to the partial loss of information arriving at the recipient if the traffic flow has some degree of flexibility. The core idea of traffic flexibilization is to reduce the sensitivity of the recipient's loss of the utility gain to partial information losses. For example, layered source coding technology can effectively enhance the flexibility of a traffic flow. The higher the flexibility of a traffic flow is, the more effective it can be applied in the above proposed flexible reservation mechanism.

4.9 Grouping Based Scheduling for Traffic Flows

In each processing block of an ITPN system, in order to improve the utilization of processing bandwidth and reduce the complexity of scheduling resources, the processing bandwidth is usually divided into certain divisions, which are then scheduled to different traffic flow groups (such strategy is called as "grouping based scheduling" in this paper). The problem to be solved in grouping based scheduling is how to effectively group traffic flows and map specific traffic groups to specific processing bandwidth. In order to effectively support the flexible reservation mechanism, the basic idea of grouping traffic flows with various performance requirements is to organize some class B traffic flows with greater flexibility and some class A traffic flows, which are difficult to be accurately predicted (that is, their arrival characteristics have larger uncertainty), into some common groups as far as possible. And, moreover, we can further put some class B traffic flows with less flexibility and some class A traffic flows,

which is much more easier to be accurately predicted, into some other common groups. It is believed that grouping based scheduling with such grouping strategy for traffic flows can improve the utilization of the shared processing resources.

4.10 Collaborative Reservation

From the general end-to-end system model of an ITPN, it can be seen that a traffic flow is processed by several cascaded processing blocks. The basic idea of collaborative reservation is that when the earlier processing block clearly knows the arrival characteristics of a specific traffic flow, it will inform the subsequent processing blocks to accurately reserve the required processing resources in advance through the signaling system between processing blocks. It is evident that for the subsequent processing blocks, the uncertainty of the arrival of the traffic flow has been eliminated in advance. Therefore, based on the collaborative reservation mechanism, the resource reservations of the subsequent processing blocks will be more accurate and efficient.

4.11 Adaptive Allocation of Processing Bandwidth

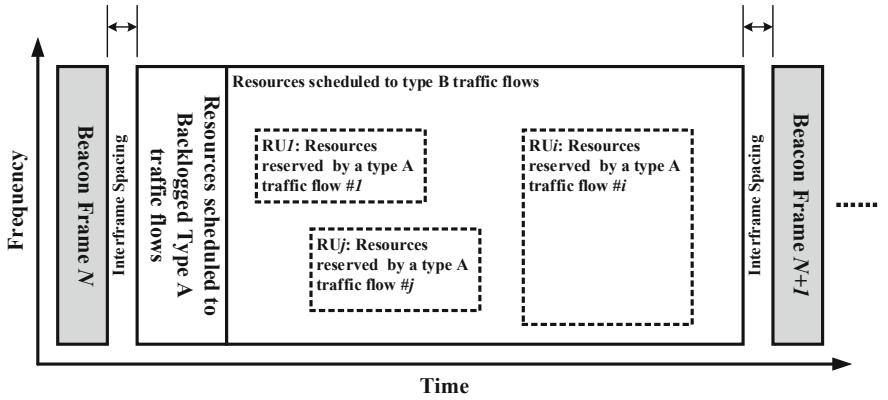
It can be seen from Eq. (7) that in order to reduce the processing delay, the corresponding processing bandwidth should be adaptively allocated according to the amount of information to be processed. That is, the larger the amount of information to be processed, the larger the processing bandwidth should be allocated.

4.12 Optimal Configuration of Processing Bandwidth

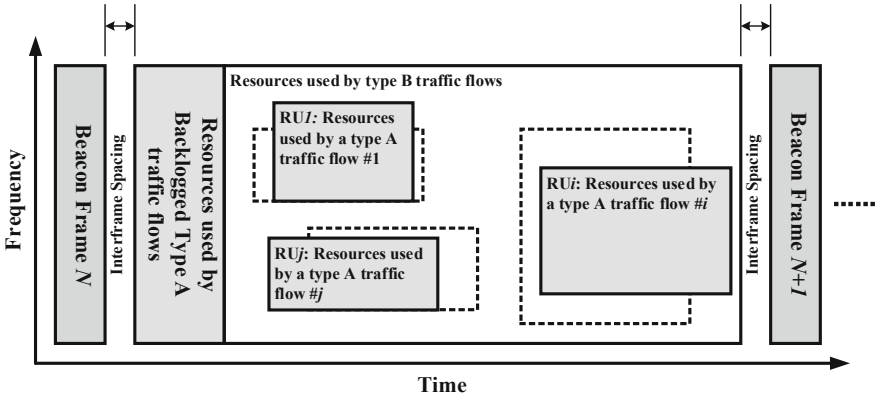
Under the given constraints, the end-to-end delay can be minimized by optimizing the allocation of processing resources among all the processing blocks in an ITPN. It is worth noting that in Eq. (8), we obtain the minimum end-to-end delay under the given total processing bandwidth of the system. In fact, considering the various requirements of the actual system, more general and complex conditions can be considered.

5 Design of a MAC Protocol Framework Supporting Ultra-Low Delay Services

In this section, based on the analysis of the end-to-end ultra-low delay of an ITPN system, we propose a general ultra-low delay MAC framework based on the basic idea of the above proposed flexible reservation. Next, we first describe the access strategies for different types of traffic flows in this MAC framework. And then, we describe a general time-frame structure that can effectively support the access strategies. It is worth noting that because of its generality, apart from IEEE 802.11 WLAN, the MAC framework proposed in this paper can be applied to various kinds of wireless communication networks.



(a) Before channel access procedure



(b) After channel access procedure

Fig. 6. The time-frame structure of the MAC framework

5.1 Basic Access Strategies of the MAC Framework

In the proposed MAC framework, we divide the traffic flows into two categories: delay sensitive traffic (class A) and delay insensitive traffic (class B).

For a class A traffic flow, the system reserves a certain amount of channel resources in an appropriate period of time according to the prediction of its arrival characteristics (i.e. traffic prediction) before its accessing into channel resources. Once the channel resource is reserved, the considered type A traffic flow has the highest access priority for the reserved resources. In other words, if the class A traffic flow needs to access into this part of resources, other traffic flows will not be allowed to use it (unless the considered class A service flow actively releases it). For a class B traffic flow, it can access the channel resources which are not reserved by using either contention-based access or scheduling-based access.

Considering the inaccuracy of traffic prediction, if some reserved resources do not have corresponding class A traffic to carry, it will release this part of reserved resources

to other class B traffic flows. On the other hand, if a class A traffic does not have enough access resources reserved for its arrived traffic, it will instantly preempt the channel resources being occupied by other class B traffic flows. The service of the preempted class B traffic flows will be suspended temporarily. In short, the exchange mechanism of channel resources between class A traffic flows and class B traffic flows is introduced in the proposed MAC framework, so as to achieve the win-win performance for both of the two types of traffic flows.

5.2 Basic Time-Frame Structure of the MAC Framework

Figure 6 shows the basic time-frame structure of the proposed MAC framework. In this time-frame structure, the access procedure of the system is divided into one superframe after another. Furthermore, each superframe can be divided into three phases: the transmission phase of beacon frame, the transmission phase of backlogged type A traffic, and the mixed transmission phase of type A and type B traffic. The following is a brief description of the three phases:

A. The transmission phase of beacon frame:
First of all, decisions for the scheduling of backlogged type A traffic, the channel reservations for type A traffic flows and the scheduling for the transmission of type B traffic flows are made. And then, the decisions made are filled into the beacon frame to be transmitted.

B. The transmission phase of backlogged type A traffic:
In this phase, it is necessary to arrange the transmission of backlogged type A traffic which has not yet been transmitted in the previous superframe. The specific actions for channel access and data transmission are executed according to the **scheduling** decision made in the "transmission phase of beacon frame".

C. The mixed transmission phase of type A and type B traffic:
This phase is executed according to the access strategies for the two types of traffic flows described above. Due to the limited space, it will not be repeated here.

6 Conclusions

This paper discusses the design methodology of ultra-low delay MAC strategies and protocols for next generation WiFi. In order to make the proposed design methodology more general, a general end-to-end transmission processing model for an ITPN is proposed. Based on it, the end-to-end delay of the system is analyzed and the expression of the minimum end-to-end delay is obtained. Based on the minimum end-to-end delay expression, we reveal three key factors that determine the end-to-end delay, namely, the number of processing blocks, the size of information blocks processed and the total processing bandwidth of the system. Furthermore, key technologies and their core ideas to realize ultra-low delay services are proposed. Finally, a general ultra-low delay MAC protocol framework based on the idea of flexible reservation is

proposed. It is believed that apart from IEEE 802.11 WLAN, the MAC framework proposed can be applied to various kinds of wireless communication networks.

Acknowledgement. This work was supported in part by the National Natural Science Foundations of China (Grant No. 61771390, No. 61871322, No. 61771392, and No. 61501373), and Science and Technology on Avionics Integration Laboratory and the Aeronautical Science Foundation of China (Grant No. 20185553035, and No. 201955053002).

References

1. IEEE 802.11ax Task Group. Project authorization request. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment: Enhancements for Extremely High Throughput (EHT), pp. 1–2 (2019)
2. Saheb, S.M., Bhattacharjee, A.K., Dharmasa, P., et al.: Enhanced hybrid coordination function controlled channel access-based adaptive scheduler for delay sensitive traffic in IEEE 802.11e networks. IET Netw. **1**(4), 281–288 (2012)
3. Pei, C., Zhao, Y., Liu, Y., et al.: Latency-based WiFi congestion control in the air for dense WiFi networks. In: 2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS), pp. 1–10. IEEE (2017)
4. Li, M., Tan, P.H., Sun, S., et al.: QoE-aware scheduling for video streaming in 802.11 n/ac-based high user density networks. In: 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), pp. 1–5. IEEE (2016)
5. Prabhu, H.V., Nagaraja, G.S.: Delay-sensitive smart polling in dense IEEE 802.11n network for quality of service. IUP J. Telecommun. **10**(1), 7–19 (2018)
6. Ahn, J., Kim, Y.Y., Kim, R.Y.: Delay oriented VR mode WLAN for efficient wireless multi-user virtual reality device. In: 2017 IEEE International Conference on Consumer Electronics (ICCE), pp. 122–123. IEEE (2017)
7. Qian, X., Wu, B., Ye, T.C.: QoS-aware A-MPDU retransmission scheme for 802.11 n/ac/ad WLANS. IEEE Commun. Lett. **21**(10), 2290–2293 (2017)
8. Zheng, H., Chen, G., Yu, L.: Video transmission over IEEE 802.11n WLAN with adaptive aggregation scheme. In: 2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–5. IEEE (2010)
9. Hajlaoui, N., Jabri, I., Taieb, M., et al.: A frame aggregation scheduler for QoS-sensitive applications in IEEE 802.11n WLANs. In: 2012 International Conference on Communications and Information Technology (ICCIT), pp. 221–226. IEEE (2012)
10. Charfi, E., Gueguen, C., Chaari, L., et al.: Dynamic frame aggregation scheduler for multimedia applications in IEEE 802.11n networks. Trans. Emerg. Telecommun. Technol. **28**(2), e2942 (2017)
11. Azhari, S.V., Gürbüz, Ö., Ercetin, O., et al.: Delay sensitive resource allocation over high speed IEEE802. 11 wireless LANs. Wireless Netw. **26**(3), 1949–1968 (2018)
12. Avdotin, E., Bankov, D., Khorov, E., et al.: Enabling massive real-time applications in IEEE 802.11 be networks. In: 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 1–6. IEEE (2019)
13. Avdotin, E., Bankov, D., Khorov, E., et al.: OFDMA resource allocation for real-time applications in IEEE 802.11 ax networks. In: 2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp. 1–3. IEEE (2019)
14. Kim, D., Yeom, I., Lee, T.J.: Mitigating tail latency in IEEE 802.11–based networks. Int. J. Commun. Syst. **31**(1), e3404 (2018)

15. Nguyen, S.H., Vu, H.L., Andrew, L.L.H.: Service differentiation without prioritization in IEEE 802.11 WLANs. IEEE Trans. Mob. Comput. **12**(10), 2076–2090 (2012)
16. Tian, G., Camtepe, S., Tian, Y.C.: A deadline-constrained 802.11 MAC protocol with QoS differentiation for soft real-time control. IEEE Trans. Ind. Inform. **12**(2), 544–554 (2016)
17. Lin, P., Chou, W.I., Lin, T.: Achieving airtime fairness of delay-sensitive applications in multirate IEEE 802.11 wireless LANs. IEEE Commun. Mag. **49**(9), 169–175 (2011)
18. Syed, I., Roh, B.: Delay analysis of IEEE 802.11e EDCA with enhanced QoS for delay sensitive applications. In: 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), pp. 1–4. IEEE (2016)
19. Wu, C., Ohzahata, S., Ji, Y., et al.: A MAC protocol for delay-sensitive VANET applications with self-learning contention scheme. In: 2014 IEEE 11th Consumer Communications and Networking Conference (CCNC), pp. 438–443. IEEE (2014)
20. Rentschler, M., Laukemann, P.: Towards a reliable parallel redundant WLAN black channel. In: 2012 9th IEEE International Workshop on Factory Communication Systems, pp. 255–264. IEEE (2012)
21. Halloush, R.D.: Transmission early-stopping scheme for anti-jamming over delay-sensitive IoT applications. IEEE Internet Things J. **6**(5), 7891–7906 (2019)
22. Pei, C., Zhao, Y., Chen, G., et al.: WiFi can be the weakest link of round trip network latency in the wild. In: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, pp. 1–9. IEEE (2016)
23. Cheng, Y., Yang, D., Zhou, H.: Det-LB: a load balancing approach in 802.11 wireless networks for industrial soft real-time applications. IEEE Access **6**, 32054–32063 (2018)
24. Choi, J., Yoo, J., Choi, S., Kim, C.: EBA: an enhancement of the IEEE 802.11 DCF via distributed reservation. IEEE Trans. Mob. Comput. **4**(4), 378–390 (2005)
25. Li, B., Tang, W., Zhou, H., et al.: m-DIBCR: MAC protocol with multiple-step distributed in-band channel reservation. IEEE Commun. Lett. **12**(1), 23–25 (2008)
26. Li, B., Li, W., Valois, F., et al.: Performance analysis of an efficient MAC protocol with multiple-step distributed in-band channel reservation. IEEE Trans. Veh. Technol. **59**(1), 368–382 (2009)
27. Singh, S., Acharya, P.A.K., Madhow, U., Belding-Royer, E.M.: Sticky CSMA/CA: implicit synchronization and real-time QoS in mesh networks. Ad Hoc Netw. **5**, 744–768 (2007)
28. Joe, I.: QoS-aware MAC with reservation for mobile ad-hoc networks. In: IEEE 60th Vehicular Technology Conference, VTC 2004-Fall (2004)
29. Sheu, S., Sheu, T.: A bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LANs. IEEE J. Sel. Areas Commun. **19**, 2065–2080 (2001)
30. Ahn, C.W., Kang, C.G., Cho, Y.Z.: Soft reservation multiple access with priority assignment (SRMA/PA): a novel MAC protocol for QoS-guaranteed integrated services in mobile ad-hoc networks. In: Vehicular Technology Conference Fall 2000, IEEE VTS Fall VTC2000, 52nd Vehicular Technology Conference (Cat. No. 00CH37152), vol. 2, pp. 942–947. IEEE (2000)
31. Jiang, S., Rao, J., He, D., et al.: A simple distributed PRMA for MANETs. IEEE Trans. Veh. Technol. **51**(2), 293–305 (2002)
32. Bankov, D., Khorov, E., Lyakhov, A., et al.: Enabling real-time applications in Wi-Fi networks. Int. J. Distrib. Sens. Netw. **15**(5), 1550147719845312 (2019)