Reservation based Resource Management for SDN-based UE Cloud

Guolin Sun[1], Dawit Kefyalew[1], Guisong Liu[1]
[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]
*Corresponding author: Guolin Sun

Received May 25, 2016; revised September 2, 2016; accepted September 27, 2016;
published December 31, 2016

Abstract

Recent years have witnessed an explosive growth of mobile devices, mobile cloud computing services offered by these devices and the remote clouds behind them. In this paper, we noticed ultra-low latency service, as a type of mobile cloud computing service, requires extremely short delay constraints. Hence, such delay-sensitive applications should be satisfied with strong QoS guarantee. Existing solutions regarding this problem have poor performance in terms of throughput. In this paper, we propose an end-to-end bandwidth resource reservation via software defined scheduling inspired by the famous SDN framework. The main contribution of this paper is the end-to-end resource reservation and flow scheduling algorithm, which always gives priority to delay sensitive flows. Simulation results confirm the advantage of the proposed solution, which improves the average throughput of ultra-low latency flows.

Keywords: software defined network, resource reservation, flow scheduling, effective bandwidth

http://dx.doi.org/10.3837/tiis.2016.12.003 **ISSN : 1976-7277**

1. Introduction

Wireless ubiquitous access is a method of allowing anywhere anytime network access by deploying many wireless technologies throughout the physical environment of the user. Cities around the world are currently considering building expensive wireless infrastructure. In urban areas, resident operated Wi-Fi Access Points (APs) are dense enough to achieve ubiquitous Internet access. In addition, a variety of wireless and mobile access networks, such as WiMAX, 3G and LTE are ubiquitously coexisting and massively deployed. These radio ecosystems cooperate to provide diverse services with various wireless network topology and different performance metrics, such as throughput, latency, energy and so on. In order to achieve ubiquitous access, one must consider different challenges such as the heterogeneous nature of the network and different Quality Of Service (QoS) requirements of the contents[1, 2]. Network resources like radio spectrum, link bandwidth and buffer space are all the fundamental basis for satisfying diverse QoS requirements of different applications. Hence, resource reservation is one of the most important components of the overall resource management process. As the next generation wireless networks cater for various kinds of applications with varying degree of QoS parameters, resource management holds a key position to optimize the overall system performance.

In recent times, there has been an increase in the number of mobile devices that access variety of services on the internet. For example, traffic related to video content accessed on the mobile devices has increased rapidly, and the trend is expected to continue. Moreover, 5G network will have strong requirements like Fiber-like user experience, less than one millisecond latency to support emergency services. At least 1 Gb/s or more data rates to support ultra-high definition video and virtual reality applications. Current solutions to solve this problem cannot provide such a strong QoS guarantee and this leads to service interruption and high latency that delay sensitive traffics. Furthermore, current methods exhibit very poor performance in terms of throughput. In near future, 5G is expected to have the set of technical components and systems needed to handle these requirements and overcome the “limits” of current systems. Internet of Things (IoT) allows different devices to communicate with each other without the intervention of human. These devices are characterized by their strong ultra-low latency requirements. The outcome of missing a tiny fraction of latency in IoT environment ranges from loss data and severe damage of device to serious injury or death of people. The decoupling of control plane and forwarding plane in SoftwareDefined Network (SDN) has proven to be effective for the high-bandwidth, dynamic nature of current and future networking technologies[3, 4].

To deal with new requirements from ultra-low latency applications, the Tatctile Internet is expected to provide services really critical aspects of society[5]. However, till now, ultra-low latency wireless transmission theory is still not well set up[6]. In this paper, we are motivated to propose a unified resource reservation and scheduling for cloud UE on the top of SDN framework. The main goal is to overcome the challenges of guaranteeing the required QoS requirement for the delay sensitive traffic. The main contribution of this paper can be summarrized as two: 1) providing a programmable resource manangement method by SDN controller to decouple delay sensitive traffic and best-effort traffic. 2) providing a resource reservation in MAC layer and a scheduling algorithm to gurantee low delay performance of delay-sensitive traffic.

The rest of the paper is organized as follows. We review the previous works in resource management and flow scheduling in Section 2. In Section 3, we describe the architecture of the proposed unified resource reservation and flow scheduling algorithm. Section 4 shows the simulation results. Finally, we conclude our work in Section 5.

2. Related Work

The medium access protocol is one of the most important aspects of any wireless network because it has direct effect on guaranteeing the QoS requirement of any flow. Despite their exceptional benefits, wireless networks have problems like unreliable wireless channel due to thermal noise, shadowing and multi-path fading. This makes assuring QoS requirements in wireless networks difficult. In addition, a good medium access protocol should deal with signaling overhead, fairness, protocol complexity, mobility, low power consumption, efficient resource utilization and different traffic classes. Hence, developing a QoS aware medium access protocol is not a trivial task.

In Wi-Fi networks, one approach to guarantee the QoS requirement of flows in wireless network is the use of different backoff related values for different stations. The main purpose of this approach is to increase the chances of winning the contention window by giving a minimum backoff value to station with high priority flow. Eventually, choosing the same backoff values for different stations leads to collusion and performance degradation. The authors in [7] proposed a solution for the above problem by using a mechanism which allows stations to broadcast their backoff related value to other stations so that there will not be a problem on choosing the same value. The Priority based QoS-Aware MAC Protocol (PQAMP) in [8] defines four different traffic classes and assigns different range of backoff values to each class. A promising improvement to the above technique is the use of different inter-frame space (IFS) value together with different backoff values for delay sensitive flows and other flows. The combination of these two techniques [9,10] has been proposed in different papers however, they cannot guarantee the strict QoS requirements of ultra-low latency flows.

Another technique to guarantee the QoS requirement of different flows is to give the transmission opportunity to a specific station. Once a station holds this opportunity, other stations stop trying to access the medium until this transmission opportunity is over. This technique can improve the system throughput of the AP but it must be used carefully since it might starve other stations. IEEE 802.11 proposed a solution for Enhanced Distributed Channel Access (EDCA) called TXOP (Transmission Opportunity), in which stations are allowed to transmit several frames continuously [11].

Advance reservation, which ranges from immediate reservation to future reservation, is proposed in different papers. Immediate reservation can be viewed as advance reservation that starts at “now” and future reservation can be viewed as advance reservation of resources for some time in the future[12]. The reserved resources can be a link bandwidth or buffer space. Early researches on advance reservation were concentrated on reservation protocols, such as RSVP[13], admission control mechanisms[14], and routing algorithm for network with advance reservations[15]. In recent Device to Device (D2D) applications, a smart resource reservation schema with channel quality detection is recommended in a frequency-time grid in terms of RBs in the 3GPP LTE system, as opposed to classical reservation Aloha where time slots are reserved over the whole bandwidth [16]. Unfortunately, there is still not any feedback link in data plane considered in the current LTE D2D communication. Recently, a reservation based collision aware resource access approach is proposed for the D2D communications [17].

The basic idea is to utilize the unused resource in data region for possible distributed coordination to avoid the collisions. Eventually, pure reservation protocols and mechanisms are not good enough to guarantee the current strong QoS requirements needed by ultra-low latency flows. In addition, till now, there is still now discussion on the scheduling problem for mixed ultra-low latency flows.

Traffic prediction is a technique used to forecast the characteristics or the direction of the UE in wireless networks. One can use this prediction to decide on how to serve the upcoming flow. The use of traffic prediction in advance reservation technique is proposed in [18][19] but this method is very complex and finding optimal solution is a challenging task. The authors in [20] proposed a scheduling-based reservation MAC protocol for wireless mesh networks. Their goal was to limit the waste of bandwidth through maximizing the slots utilization rate. As a more reliable solution, we propose the combination of the priority-based scheduling, resource reservation based access, and traffic prediction based resouce estimation in this paper.

3. The Proposed Method

In this paper, a joint resource reservation based flow scheduling algorithm is proposed. The proposed method allows end-to-end resource reservation for cloud UE, which means that the bandwidth is pre-reserved in all intermediate nodes between the sending and the receiving UEs. Most specifically, wireless mediums in the AP and bandwidth related parameters will be reserved in the intermediate routers. The proposed method is based on the famous SDN framework and it is flexible. As shown in **Fig. 1**, the SDN controller at least includes flow scheduler and route manager. It is located at the edge of access network neighbored to an edge router. The flow scheduler results from SDN controller will be mapped to each entity in the network architecture.

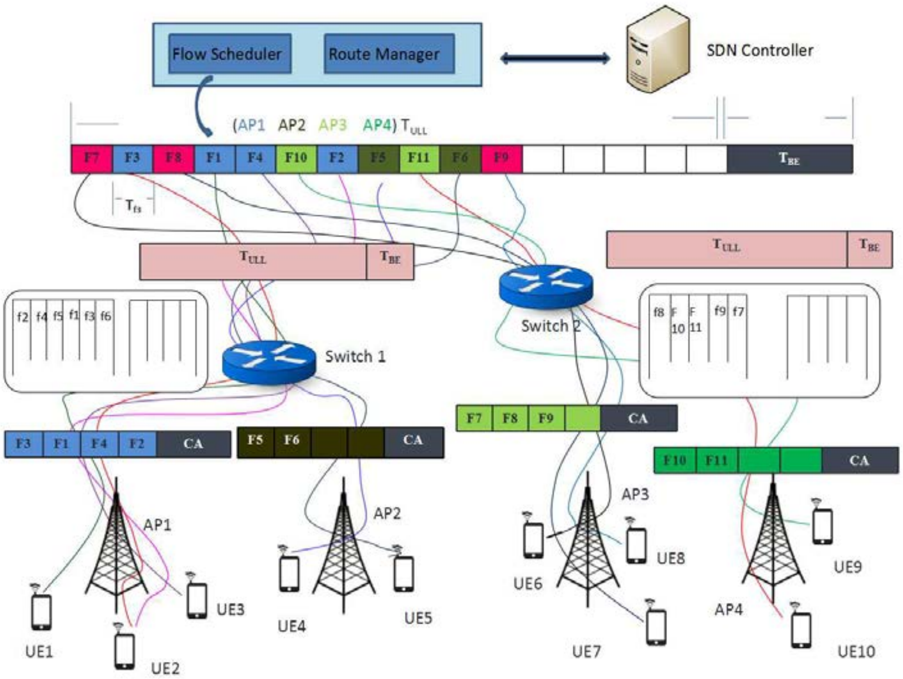


Fig. 1. Network architecture

We used the idea of mobile cloud computing to propose a cloud UE environment, which combines cloud computing, mobile computing, and wireless networks. The ultimate goal of this method is to enable execution of rich mobile applications on a plethora of UEs, with a rich user experience and different QoS requirements. The advantage of the proposed method is the UE's ability to access different services ubiquitously. The UE will access different services with different devices anywhere, anytime via wireless networks regardless of heterogeneous environments and platforms.

Fig. 1 shows the network architecture used for our work in this paper. The term "resource" in this paper is used to indicate the bandwidth required by each flow in the cloud UE. The scheduling period of each forwarding device and Access Point (AP) is divided into two sub-scheduling period, namely ultra-low latency period (T_{ULL}) and best effort period (T_{BE}) sub-scheduling frames. In T_{ULL} sub-scheduling period, delay sensitive traffics will be served and in T_{BE} sub-scheduling period, BE traffics will be served.

3.1 System Architecture

Fig. 2. shows the architecture of the proposed method. The controller is responsible for allocating resources needed to guarantee the QoS requirements of delay sensitive flows. The controller is also responsible to schedule all flows in such a way that delay sensitive flows can satisfy their strong QoS requirements. Once controller determines the resources needed to guarantee a flow, it will update the flow table of forwarding devices to enforce the reserved bandwidth. The forwarding devices use flow characteristics information and allocated bandwidth information from the controller to guarantee the QoS requirements of each flow.

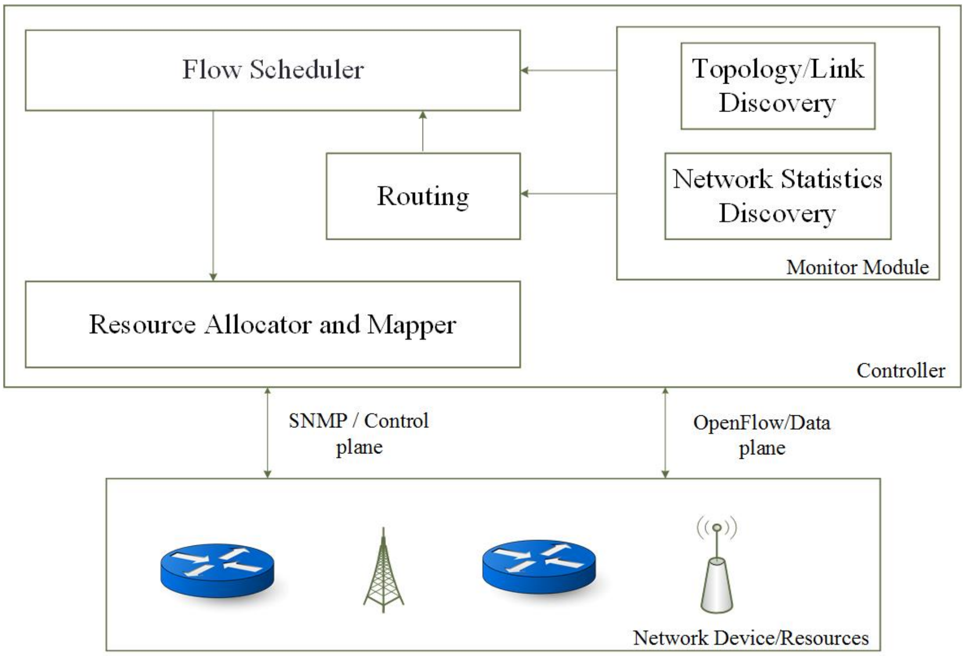


Fig. 2. System Arcitecture

The controller is composed of four separated modules namely, Monitor module, Flow Scheduler module, Routing module and Resources allocator & mapper module. The monitor module is responsible for periodically monitoring the network statistics of forwarding devices.

By communicating with forwarding devices using SNMP protocols, it's also responsible for discovering different topology information. Routing module is responsible for finding the path between the sending and the receiving UEs. Routing module can be implemented using simple Dijkstra's shortest path algorithm to find the paths. Flow scheduler module is responsible for scheduling incoming delay-sensitive flow by running the proposed flow scheduling algorithm. This module uses the path information of the requesting flow from routing module and network statistics information from monitoring module to schedule flows. Finally, resources allocator and mapper module allocates the resources for requesting flows and map the results to the each forwarding devices throughout the path.

3.2 Resource reservation

The proposed MAC protocol is composed of two different access methods: a contention free access (CFA) and contention access (CA). CFA intends for the transmission of delay sensitive traffics and CA intends for the transmission of best-effort traffics. The CFA method is also used to send resource reservation requests (RRR). CFA period is divided into periodic Time Division Multiple Access (TDMA) frames, each consisting of F time-slots, where each time-slot supports a fixed maximum throughput per time-slot. The legacy CSMA/CA is used to access the channel in the CA method. This protocol requires the UE to perform time synchronization with the central controller before sending RRR. The controller maintains this synchronization for the UE across all APs regardless of the UE movement. The reason behind this is to eliminate the complex process of time synchronization performed by the UE every time it moves to a different AP. The frame structure of proposed MAC protocol is shown in **Fig. 3**. We divide the time axis into a series of fixed-length periods called frame, in each frame a CFA followed by a CA. We divide CFA further into fixed time-slots. As mentioned before, CA is only used for time-slots reservation and for transmissions of BE traffic. And then the multiple data packets are transmitted during the CFA.

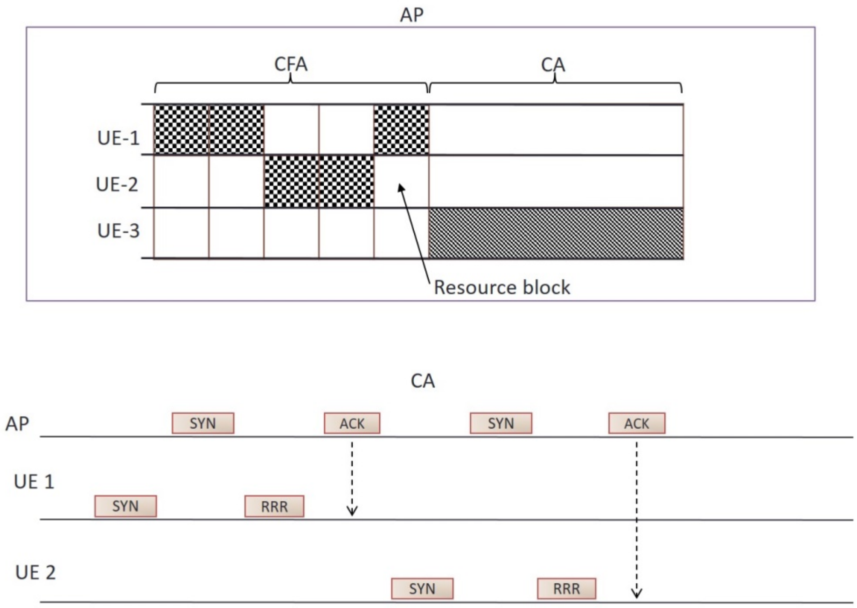


Fig. 3. Resource reservation method

All UEs with traffic requirements reserve multiple periodic time-slots for its data packets transmissions with only one successful handshake in CA. Frankly speaking, each UE which has the data transmission request is involved in contentions according to IEEE 802.11 DCF scheme, and then reserves multiple periodic time-slots in advance through only one successful resource reservation handshake in CA. Once reserved successfully, the wining UE starts to transmit the data packets in the multiple reserved time-slots during CFA and will not contend again until the next CA. Furthermore, the failed UEs keep silent during the current CFA and will re-contend in the next CA. Therefore, the proposed protocol reduces collisions in CA and provides the collision-free data transmissions in CFA. In addition, it also improves the slot usage efficiency and reduces the extra overheads. The last but not least, it provides QoS guarantee in the process of data transmission.

In the proposed MAC protocol, there are two kinds of control packets, Resource Reservation Request (RRR) packet and Acknowledgements (ACK) packet for reservation confirmation and for broadcasting reservation information to other UEs. Compared with the traditional RTS packet in IEEE 802.11, the RRR packet; add the new content, the traffic specification (TSPEC) of a flow. ACK packet holds reserved time-slot IDs, total number of reserved time-slots and time-slot usage list. Additionally, reservation table is maintained and updated at each UE through RRR-ACK negotiations, which includes the current time-slots usage condition.

The detail operation of proposed MAC protocol is divided into two processes: time-slot reservation using three-phase handshake in CA, and multi-step data transmissions with ACK responding in CFA. In the Request Phase, the UE with the delay-sensitive traffic transmission desire makes its request for reservation. In this phase, supposing UE wants to make a reservation, it firstly inspects its current reservation table to confirm whether it contains available time-slots. If there is free time-slots exist, then UE generates the TSPEC values and sends an RRR packet to the AP. In the Confirm and Broadcast Phase, the AP confirms the request of the UE and broadcasts the results if it’s valid. Upon receiving this RRR, the AP forwards this request to the controller. The SDN controller has the global view and it is responsible for managing, reserving and monitoring the resource in the network. The controller calculates the effective bandwidth needed by the flow to guarantee its QoS requirement.

Then the controller sends the total numbers of time-slots needed by the flow to AP. Henceforth, the AP decides which time-slots can be assigned to the UE for data transmissions, and then updates its time-slot usage list. Otherwise, if AP cannot response this RRR packet, UE re-sends RRR packet. Supposing there are enough available time-slots in CFA for requesting UE, the AP responses an ACK packet to UE according to RRR packet. The neighbors of the requesting UE listens the ACK packet and know that which data slots have been selected. They update their reservation table based on the ACK packet. The UE re-sends RRR packet during next CA on condition that there are no available time-slots. ACK packet transmission also informs other UEs which are in the vicinity of requesting UE, these neighbor UE label the selected time-slots as busy. And then these neighbors update their reservation table based on the ACK packet to obtain the current time-slots usage condition.

During CFA, all UEs send data packets depending on the successful reserved multiple time-slots. The other UEs, which do not reserve time-slots in current CFA keep listening until the next CA for re-contention, consequently, this protocol improves the efficiency of data packets transmission. It is worth mentioning that the proposed MAC protocol has some elements that are similar to other existing MAC protocols. (E.g., the three-phase dialogue is similar to RTS/CTS handshake in DCF). The major difference however is that the proposed MAC protocol is scheduling-based and it considers the effective bandwidth of each flow. The proposed scheduler will guarantee the adjacent packets of each flow in a good order based on an inter-ideal departure time. Besides the QoS guarantee, it also increases throughput, supports flow-level bandwidth provisioning and removes extra overhead using the less control packets negotiation.

3.3. Flow Scheduling algorithm

Consider a flow defined by a traffic specification with peak rate, the maximum packet size, the sustainable rate, and the maximum burst size, TSPEC (*h, M, r, b*). Network Calculus [21] provides us the following definitions to calculate the effective bandwidth (*eB*) of a flow.

Definition 1 (Arrival Curve): Given a wide-sense increasing function α defined for t ≥ 0 we say that a flow U is constrained by α if and only if for all s ≤ t:

$$U(t) - U(s) \leq \ \alpha(t-s) \tag{1}$$

We say that R has α as an arrival curve.

Definition 2 (Service Curve): Consider a system S and a flow through S with input and output function U and U*. We say that S offers a service curve β to the flow if and only if β is wide sense increasing β(0) = 0 and

$$U^{*}(t) \geq (U \otimes \ \beta)\ (t) \tag{2}$$

where operator ⊗ denotes the convolution operation in min-plus algebra defined as x ⊗ y(t) = infs:0≤s≤t {x(t − s) + y(s)}.

Definition 3 (Effective Bandwidth): Consider a system S and a flow through S with an arrival function α, for a fixed but arbitrary delay *d*, we define the effective bandwidth eB(α) of the flow as the required bit rate at s to serve the flow in a work conserving manner to guarantee the maximum delay *d*; that is

$$eB(\alpha) = \sup_{s\geq 0}\{\alpha(s)/(s+d)\} \tag{3}$$

For a flow with T-SPEC (h,M, r, b), Definition 3 can be simplifided in equation 4.

$$eB(f)\ = MAX\left\{\frac{M}{d}, r, h\left(1 - \frac{d - \frac{M}{h}}{\frac{b-M}{h-r} + d}\right)\right\} \tag{4}$$

where d is the delay requirement of the flow.

The intuition idea of our proposed flow scheduler is to schedule the packet leaving different senders in a good order to keep the time delay between two adjacent packets for each flow suitable for its QoS requirement. The ideal departure time $D^{*}(i,f)$ is used to identify the departure time for packet *i* for flow *f*. $\tau^{*}(f)$ denote the Ideal Inter-Departure Time (IIDT) between adjacent packets in a perfectly scheduled flow with zero jitter. However, in the various scenarios with multiple delay sensitive flows, the ideal departure time of each flow is limited to the slots available and it cannot reach always. Therefore, another variable, Virtual Finishing Time (VFT), is calculated and replaces the ideal departure time $D^{*}(i,f)$. The specific descript of this algorithm is explained below with the concept of effective bandwidth introduced before.

As the section 3.2 and **Fig. 3** explained, the key idea of the proposed reservation based resource allocation method is to divide the time into frames. The TDMA-based CFA and the CA are the two components used to transmit delay sensitive traffic and the best effort traffic, separately. However, the overall target of the proposed method tries to guarantee the QoS of delay sensitive flows with a higher priority and maybe sacrifice the throughput of best-effort flows. The number of slots is required to calculate for CFA and given as the sum of equivalent bandwidth of each flow. Therefore, in this scheduling algorithm, we did not include the discussion about the impact on the best effort flows in CA. Here, we assume the configured CFA period is enough and we will discuss the utilization efficiency for different schedulers.

Let the time axis be divided into scheduling frames, each consisting of F time-slots, where each time-slot supports the maximum throughput of $T_{throughput}$. Let R(f) denote the number of time-slot reservations per scheduling frame needed to support the flow, where 0≤ R(f) ≤ F. In another word, the R(f) is an identifier for delay requirement of each flow. We can calculate R(f) using the following equation.

$$R(f) = \frac{eB(f)}{floor(\frac{T_{throughput}}{M}) * M} * F \qquad (5)$$

Generalized Processor Sharing (GPS) concept in[22] presents a theory for computing packet departure times. Let $p(i,f), D(i,f)\ and\ D^*(i,f)$ denotes packet i of flow f, the actual departure time of packet $p(i,f)$ and the ideal departure time of packet $p(i,f)$ in a perfectly scheduled zero-jitter flow respectively. Let $\tau^*(f)$ denotes the Ideal Inter-Departure Time (IIDT) between adjacent packets in a perfectly scheduled flow with zero jitter.

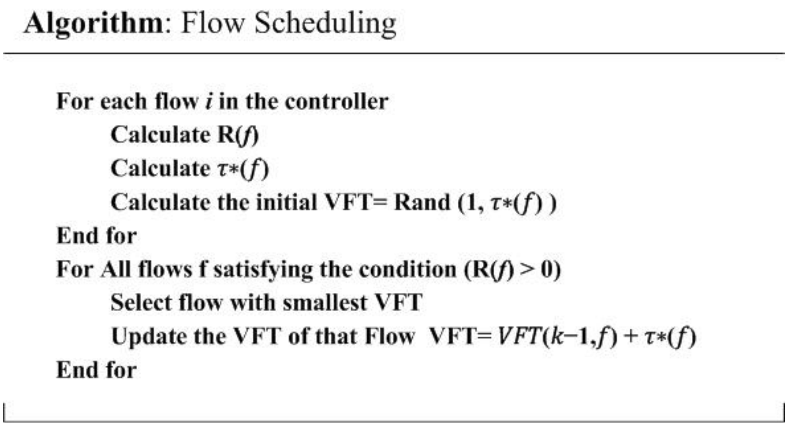
$$\tau^*(f) = \frac{F}{eB(f)} \qquad (6)$$

Therefore, we can calculate the ideal departure time of packet using the following equation.

$$D^*(i,f) = D^*(i-1,f) + \tau^*(\mathrm{f}) \qquad for\ i > 1 \qquad (7)$$

$$D^*(i,f) = \mathrm{Rand}(1, \tau^*(f)) \qquad for\ i = 1 \qquad (8)$$

where rand(*) is a random function for uniform distribution.



```
Algorithm: Flow Scheduling

For each flow $i$ in the controller
    Calculate R($f$)
    Calculate $\tau*(f)$
    Calculate the initial VFT= Rand (1, $\tau*(f)$ )
End for
For All flows f satisfying the condition (R($f$) > 0)
    Select flow with smallest VFT
    Update the VFT of that Flow  VFT= $VFT(k-1,f) + \tau*(f)$
End for
```

Fig. 4. The proposed flow scheduling algorithm

Let $\mathrm{VFT}(i,f)$ denote the virtual finishing time of $p(i,f)$. With the following equation (9) and (10), we compute the VFTs for packet i in flow f, which are used to schedule all delay sensitive flows.

$$VFT(i,f) = VFT(i-1,f) + \tau^*(\mathrm{f}) \qquad for\ \ i > 1 \qquad (9)$$

$$VFT(i,f) = \mathrm{Rand}(1, \tau^*(f)) \qquad for\ \ i = 1 \qquad (10)$$

The proposed algorithm uses the VFT of each packet in a flow to give priority to delay sensitive flows. The proposed scheduling algorithm starts to work by firstly calculating $R(f), \tau^*(f)$ and initial VFT values. Then it will select a flow with minimum VFT value for service. The VFT value of the selected flow is updated by adding IIDT value of that flow to its current VFT value. **Fig. 4** shows the proposed flow scheduling algorithm.

The proposed flow scheduling algorithm is compared with Weighted Fair Queuing (WFQ) [23] and Round-robin scheduling algorithms. WFQ is a technique used in packet-switched networks to guarantee the bandwidth requirement of different flows. WFQ is both a packet based implementation of the GPS policy, and a natural generalization of fair queuing (FQ). The main goal of WFQ is to let all flows share the limited total bandwidth in a way that a flow with maximum weight gets the maximum portion of the total bandwidth. In WFQ, each flow will be configured with specific weight w_i and then flow i will achieve an average bandwidth of B(i). The network administrator is responsible for assigning weights to each flow.

Round-robin is another scheduling technique used in packet-switched networks to give each flow a service opportunity of equal portions in circular fashion. This scheduling algorithm is well known for being simple and starvation free. Round-robin scheduling algorithm is very easy to implement and it handles all flows with same priority. Round-robin scheduling can also be used in centralized wireless network where different nodes share a single frequency channel. Base stations can use this algorithm to reserve a time-slots for mobile nodes in circular order and provide fairness.

4. Simulation Results and Analysis

In this section, we present the numerical results and the performance evaluation of our work. Five UEs associated with single AP is considered and each UE will generate one flow with different TSPEC parameters. Generic Cell Rate Algorithm (GCRA) [24] is used to shape the traffic of each flows according to the TSPEC parameters. To analyze the performance of the proposed scheduling algorithm, different AP capacities are used. To be specific, this simulation is conducted with AP bandwidth of 1Mbps, 1.5 Mbps and 2Mbps.

Table 1 summarizes the key parameter configurations used in this simulation. The scheduling period of the AP is divided into 50 equal sized time-slots. Each time-slot supports a maximum throughput of 20kb per time-slot for AP with a bandwidth of 1Mbps, 30kb per time-slot for AP with a bandwidth of 1.5Mbps and 40kb per time-slot for AP with a bandwidth of 2Mbps. The reason behind using these three AP bandwidths (1Mbps, 1.5 Mbps and 2Mbps) for different simulation setup is, to evaluate the performance those three scheduling algorithms (proposed, WFQ and round-robin) in the environment where there is limited bandwidth, enough bandwidth and excesses bandwidth.

Table 1. System configuration

Simulation Parameter	Value
Number of APs	1
Number of UE	5
Number of flow per UE	1
AP's Bandwidth	1Mbps, 1.5Mbps and 2Mbps
Number of time-slots (K)	50
Throughput	20Kb, 30Kb and 40Kb per Time-slot

A. Scenario configuration

In the equation (4), it is stated that the effective bandwidth (eB) of a flow is the maximum of the three parameters. Since the proposed flow scheduling algorithm is based on eB, this simulation used two types of scenario configuration. The first configuration contains five flows with their eB equal to the first parameter of equation (4). Similarly, the second configuration contains five flows but, their eB is equal to the second parameter of equation (4). In other words, the above two scenario configuration can mean scenario configuration for delay sensitive flows and scenario configuration for flows with high packet generation rate.

The aim of the first configuration is to evaluate the performance of the three scheduling algorithm when there is a need to guarantee the delay sensitive flows. The aim of the second configuration is to evaluate the performance of the three scheduling algorithm when there is a need to guarantee flows with high buffer requirements. The TSPEC parameters of each flows generated randomly to make the eB comply with the goal of each scenario configuration. The two scenario configurations are shown in **Table 2** and **Table 3**.

Table 2. The first scenario configuration

Flow	**h**	**M**	**r**	**b**	**Delay**	**W**	**eB**
	(Kbps)	(Kb)	(Kbps)	(Kb)	(Sec)	(WFQ only)	(Kbps)
1	6000	9	160	80	0.05	4	180
2	3400	4	360	152	0.01	6	400
3	1260	12	108	37	0.1	3	120
4	5640	10	40	100	0.2	2	50
5	2160	12	576	60	0.02	5	600

Table 3. The second scenario configuration

Flow	**h**	**M**	**r**	**B**	**Delay**	**W**	**eB**
	(Kbps)	(Kb)	(Kbps)	(Kb)	(Sec)	(WFQ only)	(Kbps)
1	6000	12	60	80	0.5	3	60
2	3400	2	90	152	0.1	4	90
3	1260	8	900	37	0.5	6	900
4	5640	10	600	100	0.2	5	600
5	2160	11	90	60	0.3	4	90

Note that the 'W' column in the above two tables is to show that the weight (W) needed for WFQ scheduling algorithm. Each weight in the two scenario configurations is generated according to the behavior of each flow and the nature of the scenario configuration.

B. Simulation results

Before the simulation results are introduced, two related metrics are defined to evaluate the performance of the proposed algorithm. One is the average throughput of each flow in two different scenarios. The average throughput actually show how many packets have been sent out during one schedule period. The time-slot utilization is another metric to show the average utilization of each time slot during a whole CFA. As shown in the **Fig. 5**, the color bar in black is the occupied time duration in each slot. The color bar in white or empty is the unoccupied time duration. The average time-slot utilization can be gotten by the sum of occupied 'black'duration over the sum of the assigned time slots. Another indirect metric can be the number of assigned time slots in a whole CFA period. The number assigned time slots can provide us the information of resource efficiency which algorithm performs better.

Fig. 5. shows the average throughput of all five flows from the first scenario configuration. The figure illustrates the result of each three scheduling algorithms in terms of average throughput in Kbps.

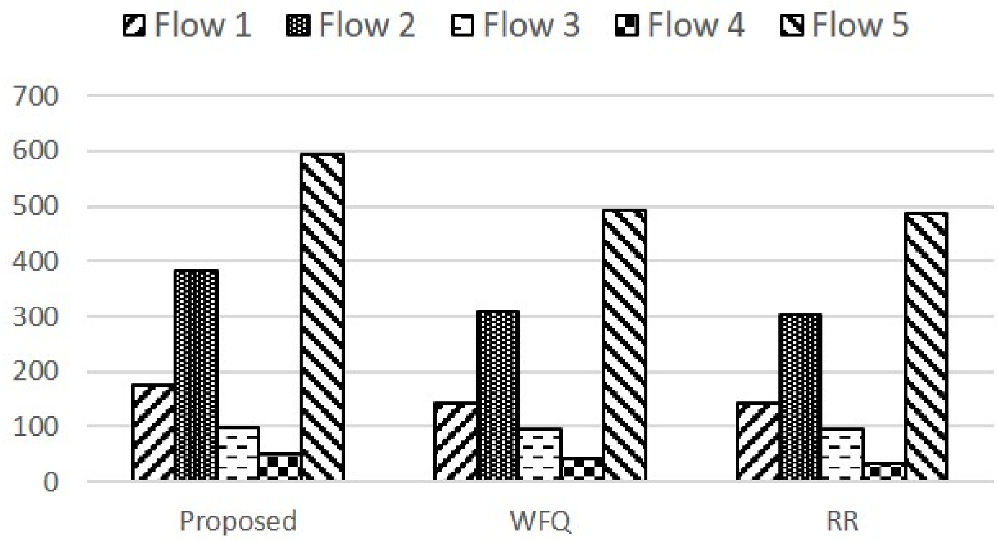


Fig. 5. The average throughput of the first scenario configuration

As you can see from **Fig. 5**., the proposed algorithm has better average throughput for flow 1, flow 2 and flow 5 (with delay requirements of 50ms, 10ms and 20ms respectively), whereas for other flows the proposed algorithm has slightly equal average throughput. Notice that the sustainable rate of flow 3 and flow 4 is 108Kbps and 40Kbps respectively. This results come from this sceniaro configuration, which provides the five flows superflous enough time slots, as shown in the **Fig. 6**. The number of assigned time slots are 38 of 50 used for all of the three algorithms and others are free.

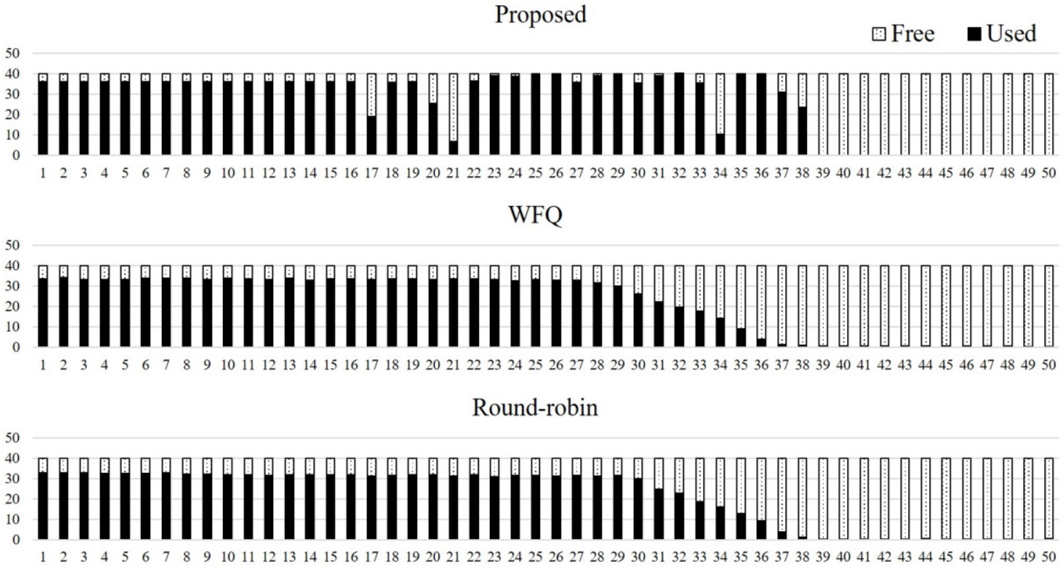


Fig. 6. The time-slot utilization of the first scenario configuration

Fig. 6. shows the average time-slot utilization of each scheduling algorithm. The figure indicates that the proposed algorithm has better average time-slot utilization for AP with

2Mbps bandwidth. Keep in mind that the maximum achievable throughput of this configuration is 40Kb per time-slot.

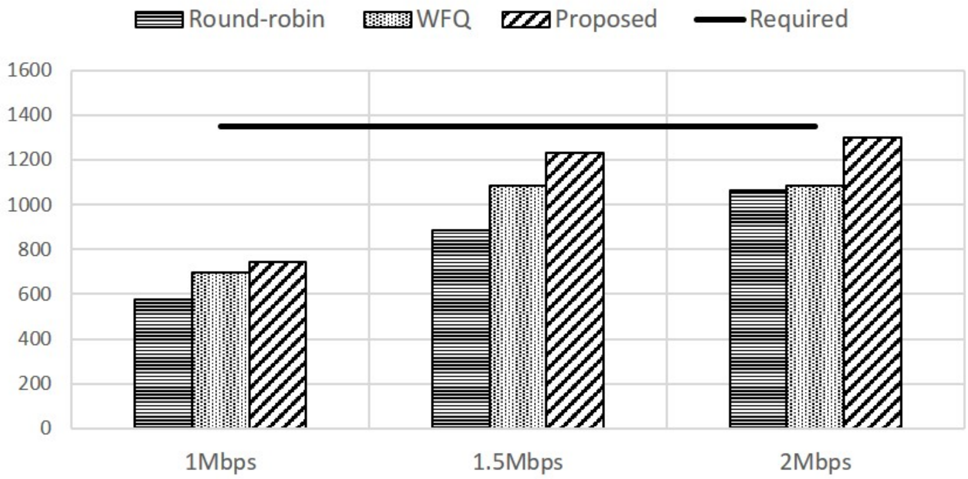


Fig. 7. Summery of system throughput comparison for the first scenario configuration

Fig. 7. shows the system throughput comparison between the three AP capacities for the first scenario configuration with 3 different bandwidth limitation. For this scenario, with the equivalent bandwidth above 1.5Mbps is allocated to CFA period, the proposed scheduling algorithm can achieve 1200Kbps throughput, which is the closest to the required throughput. This figure illustrates that the proposed scheduling algorithm has better system throughput in each AP capacitiy. **Fig. 8**. shows the average throughput of all five flows from the second scenario configuration. The figure illustrates the result of each three scheduling algorithms in terms of average throughput. As you can see from the figure, the proposed algorithm has better average throughput for flow 3 and flow 4 (they has huge buffer requirement 900Kbps and 600Kbps respectively), whereas for other flows the proposed algorithm has slightly equal average throughput. Notice that the sustainable rate of flow 1, flow 2 and flow 5 is 60Kbps, 90Kbps and 60Kbps respectively.

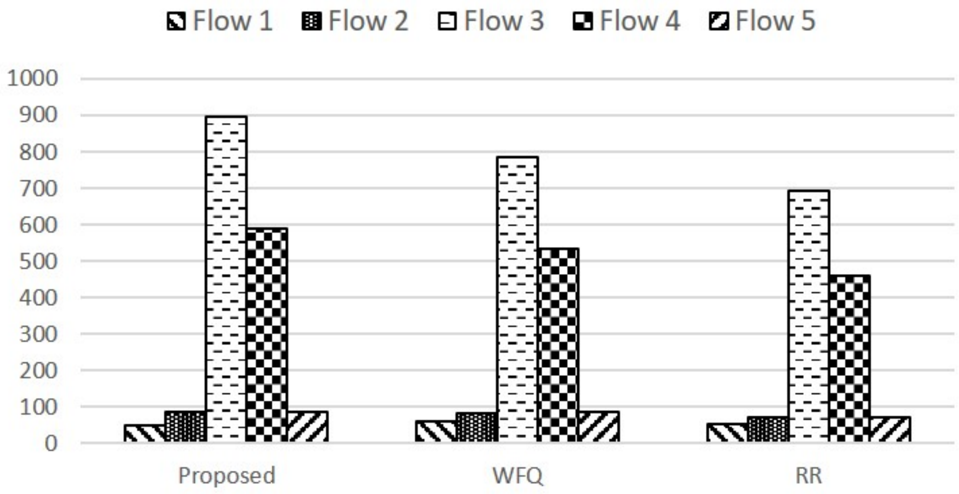


Fig. 8. The average throughput of second scenario configuration

Fig. 9 shows the time-slot utilization of each three scheduling algorithms. The figure indicates that the proposed algorithm has better time-slot utilization for AP with 2Mbps bandwidth. Keep in mind that the maximum achievable throughput of this configuration is 40Kb per time-slot.

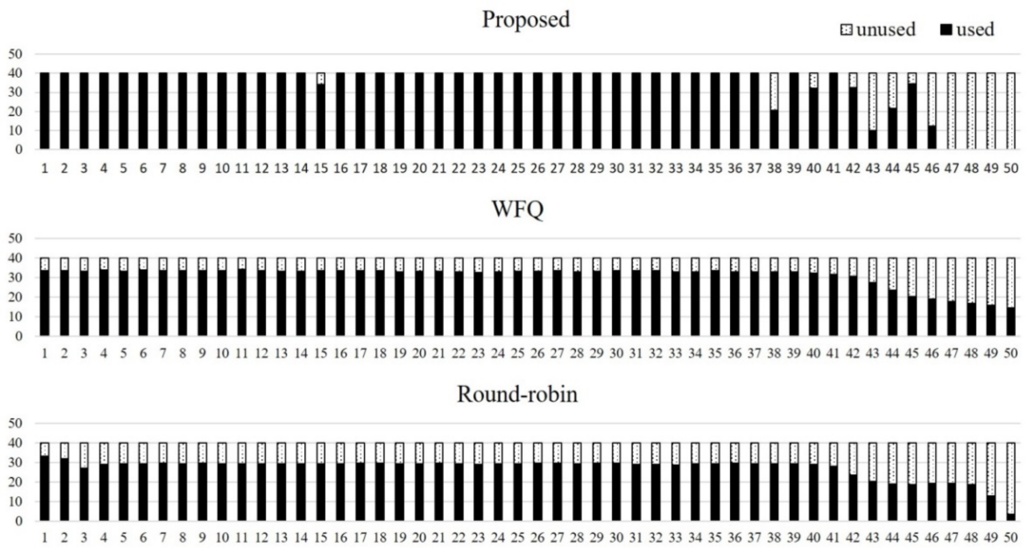


Fig. 9. The time-slot utilization of the second scenario configuration

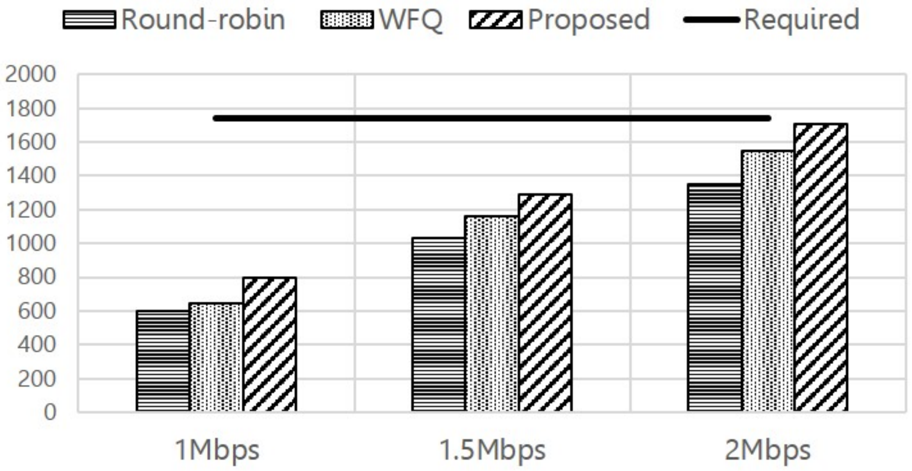


Fig. 10. Summery of system throughput comparison for the second scenario configuration

Fig. 11 shows the system throughput comparison between the three AP capacities for the second scenario configuration with 3 different bandwidth limitation. For this scenario, with the equivalent bandwidth above 2Mbps is allocated to CFA period, the proposed scheduling algorithm can achieve 1600Kbps throughput, which is the closest to the required throughput. This figure illustrates that the proposed scheduling algorithm has better system throughput in each AP capacities.

From the above two simulation results we realize that

- Round-robin scheduling algorithm gives priority for flow with highest packet size M
- WFQ scheduling algorithm give priority for flow with highest sustainable rate r and

- The proposed scheduling algorithm give priority for flow with highest *eB*.

Hence we conclude that the proposed algorithm outperforms the other two scheduling algorithms for delay sensitive flows since it considers the delay requirement the flow to give priority. Safety-critical traffic such as emergency communications are always characterized by having very-low frequency of occurrence, extremely short delay constraints, and small packet size. The proposed algorithm best satisfies the requirement of such traffics as compared to the other two scheduling algorithms.

5. Conclusion

In this paper, we proposed a joint resource reservation and flow scheduling algorithm for cloud UE. Our proposed method used the power of SDN to reserve and schedule resources for delay sensitive flows. The basic idea of our proposed method is to abstract a collection of UEs as cloud UE and reserve resources to each flow based on the result of the proposed flow scheduling algorithm. A major benefit of our method is; it does not solely consider the maximum packet size or the rate of the flow; rather it is based on the effective bandwidth of the flow. Hence, it is very applicable for safety-critical and emergency traffics. Simulation results show that the proposed method has capabilities to satisfy the QoS requirements of delay sensitive flows. The results also show the proposed method has better average throughputs for delay sensitive flows. For the future work, we planned to improve the system throughput each AP by introducing bandwidth-borrowing schemes. The ultimate goal is to borrow bandwidth from another AP when there is a limited amount of bandwidth left to satisfy the QoS requirements of a flow.

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] Jo M, Maksymyuk T, Strykhalyuk B, et al., "Device-to-Device Based Heterogeneous Radio Access Network Architecture for Mobile Cloud Computing," *IEEE Wireless Communications*, Vol.22, No.3 , pp. 50-58, June 2015. Article (CrossRef Link)

[2] Jo M, Maksymyuk T, Batista R L, et al., "A Survey of Converging Solutions for Heterogeneous Mobile Networks," *IEEE Wirelees Communicatons*, Vol 21, No 8, pp.54-62, Dec. 2014. Article (CrossRef Link)

[3] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," in *Proc. of IEEE Access*, vol. 3, no., pp. 1206-1232, 2015. Article (CrossRef Link)

[4] F. Hu, Q. Hao and K. Bao, "A Survey on Software-Defined Network and OpenFlow: From Concept to Implementation," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2181-2206, Fourthquarter 2014. Article (CrossRef Link)

[5] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultra-reliable,and low-latency wireless: The art of sending short packets," in *Proc. of the IEEE*, August 2016. Article (CrossRef Link)

[6] Simsek M, Aijaz A, Dohler M, et al., "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, 34(3): 1-1, 2016. Article (CrossRef Link)

[7] R. Baldwin, N. Davis IV and S. Midkiff, "A real-time medium access control protocol for ad hoc wireless local area networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 3, pp. 20-27, 1999. Article (CrossRef Link)

[8] Nityananda Sarma and Sukumar Nandi, "A priority based QoS-Aware MAC protocol (PQAMP) in mobile ad hoc networks," in *Proc. of the 4th ACM symposium on QoS and security for wireless and mobile networks (Q2SWinet '08)* ACM, New York, NY, USA, 79-82, 2008. Article (CrossRef Link)

[9] Yang Xiao and Haizhon Li, "Local data control and admission control for QoS support in wireless ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1558-1572, Sept. 2004. Article (CrossRef Link)

[10] A. Ksentini, A. Gueroui and M. Naimi, "A new IEEE 802.11 MAC protocol with admission control for sensitive multimedia applications," in *Proc. of IEEE Global Telecommunications Conference*, 2005., St. Louis, MO, pp. 5 pp.-3006, 2005. Article (CrossRef Link)

[11] IEEE Draft Standard for Information Technology-Telecommunications and Information exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment : Radio Resource Measurement of Wireless LANs (Amendment to IEEE Std 802.11-2007)," in *Proc. of IEEE Active Unapproved Draft Std P802.11k_D8.0*, May 2007 , vol., no., pp., 2007. Article(CrossRef Link)

[12] Z. Meng, W. Chanle and Y. Gang, "Dynamic Advance Reservation for Network Resource," *Intelligent System Design and Engineering Applications (ISDEA)*, in *Proc. of 2013 Third International Conference on*, Hong Kong, pp. 618-621, 2013. Article (CrossRef Link)

[13] Schill, A., S. Kühn, and F. Breiter, "Design and evaluation of an advance reservation protocol on top of RSVP," *Springer in Broadband Communications, Boston, MA*, pp. 23-40, 1998. Article(CrossRef Link)

[14] Ferrari, D., A. Gupta, and G. Ventre, "Distributed advance reservation of real-time connections," *Springer Network and Operating Systems Support for Digital Audio and Video*, pp. 16-27, 1995. Article(CrossRef Link)

[15] R. A. Guerin and A. Orda, "Networks with advance reservations: the routing perspective, " in *Proc. of IEEE Nineteenth Conference of the IEEE Computer and Communications Societies INFOCOM 2000, Tel Aviv*, pp. 118-127 , 2000. Article(CrossRef Link)

[16] Yaacoub, E., "On real-time smart meter reading using OFDMA based random access," in *Proc. of 17th IEEE Mediterranean Electro-technical Conference*, pp. 156 – 162, 2014. Article(CrossRef Link)

[17] 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)

[18] S. Malarkkan and V. C. Ravichandran, "Performance analysis of mobility based predictive call admission control with resource reservation in WCDMA cellular systems," in *Proc. of 2005 Asia-Pacific Conference on Communications, Perth, WA*,pp. 154-158 , 2005. Article(CrossRef Link)

[19] R. K. Sharma, P. Kamal and S. P. Singh, "A latency reduction mechanism for virtual machine resource allocation in delay sensitive cloud service," *Green Computing and Internet of Things (ICGCIoT),* in *Proc. of 2015 International Conference on, Noida,* pp. 371-375, 2015. Article(CrossRef Link)

[20] G. Boudour, C. Teyssié and Z. Mammeri, "Scheduling-based reservation MAC protocol for bandwidth and delay optimization in wireless mesh networks," in *Proc. of 2008 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Avignon,* pp. 272-277, 2008. Article(CrossRef Link)

[21] Le Boudec, J.-Y. and P. Thiran, "Network calculus: a theory of deterministic queuing systems for the internet, " *Springer Science & Business Media*, Vol. 2050, 2001. Article(CrossRef Link)

[22] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple node case," *IEEE/ACM Transactions on Networking*, vol. 2, no. 2, pp. 137-150, Apr., 1994. Article(CrossRef Link)

[23] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344-357, Jun 1993. Article(CrossRef Link)

[24] Perros, H.G., "An introduction to ATM networks, " *John Wiley & Sons*, 2002. 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, wireless networks.

Dawit Kefyalew received his BEng in software engineering from Adama Science and Technology University, Adama, Ethiopia, in 2013. From 2013 to 2014, he worked as a graduate assistance for Dilla University. He is currently pursuing his MS on Computer Science in the University of Electronics Science and Technology of China (UESTC), Chengdu, Sichuan, China. His interests include cloud computing, mobile communications, and software defined networking (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.