

Joint Resource Reservation and Flow Scheduling for Ultra-low Latency Transmission

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

Abstract—In recent times, there has been an increase in the number of mobile devices to access a variety of services on radio access network, and the trend is expected to continue. In addition, ultra-low latency services require much bandwidth and often characterized by having extremely short delay constraints. Hence, satisfying required strong QoS requirement becomes challenging task. Existing scheduling methods to solve this problem exhibit very poor performance in terms of transmission latency. In this paper, a scheduling-based resource reservation mechanism is proposed for cloud UE. Unlike other methods, the proposed algorithm in this paper considers various traffic parameters to calculate the effective bandwidth of the flow and always gives priority to delay sensitive flows under a software defined network framework. Simulation results show that the proposed scheduling algorithm improves the average throughput of ultra-low latency flows.

Index Terms—flow scheduling, resource reservation, software defined networks, effective bandwidth

I. INTRODUCTION

In order to achieve ubiquitous access, one must consider different challenges such as the heterogeneous nature of the network and different QoS/QoE requirements of the contents. Network resources like radio spectrum, link bandwidth and buffer space are the fundamental basis for satisfying diverse Quality of service (QoS) requirements of different applications. 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. Hence, scheduling and resource reservation is one of the most important components of the overall resource management process.

In recent times, there has been an increase in the number of mobile devices to access a variety of services on the internet. For example, traffic related to video content accessed by the mobile devices has increased rapidly. Moreover, 5G cellular network will have strong requirements like Fiber-like user experience, less than one millisecond end-to-end latency to support emergency services, at least 1 Gb/s or more data rates to support ultra-high definition video and virtual reality applications[1][2]. As the future data traffic is expected to be enormous, satisfying the required strong QoS requirement is become challenging task. Current existing methods to solve this problem don't guarantee the QoS requirements and lead to service interruption and high latency to delay sensitive traffic.

Advance reservation scheme which ranges from immediate reservation to future reservation is proposed in different papers [3]. The reserved resources can be link bandwidth or buffer space. Those reservation protocols and mechanisms are not good enough to guarantee the current strong QoS requirements needed by ultra-low latency flows in 5G. A scheduling-based reservation MAC protocol for wireless mesh networks was proposed [4]. Their goal was to limit the waste of bandwidth through maximizing the slots utilization rate. Weighted Fair Queuing (WFQ) and Round-robin are two popular packet scheduling algorithm in packet switch networks. The WFQ is a natural generalization of fair queuing technique in packet-switched networks to guarantee the bandwidth requirement of different flows. The main goal of WFQ is to let all flows share the limited total bandwidth in a way that a flow with the different weights. The flow with the largest weight will get the maximum portion of the total bandwidth. The network administrator is responsible for assigning weight to each flow. Round-robin is a simple scheduling technique used in packet-switched networks to give each flow a service opportunity of equal portions in circular fashion. The Generalized Processor Sharing policy (GPS) concept presents a theory for computing packet departure times, which is extended to latency sensitive flow scheduling in this paper [5].

The decoupling of control plane and forwarding plane in Software-Defined Networking (SDN) is proven to be effective for the high-bandwidth, dynamic nature of current and future networking technologies[6]. Hence, this paper is 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 delay sensitive traffics. Firstly, the proposed scheduling method could be software defined with a controller. This feature will allow it can decouple scheduling modules for different air-interface network slices, such as UE slice and ultra-low latency IoT slice. Secondly, the scheduling based resource reservation is on-demand based on the traffic analysis instead of static weight in WFQ method. Hence, it is adaptive to traffic and network.

The rest of the paper is organized as follows. In Section II, we describe the system architecture, protocol and the proposed unified resource reservation and flow scheduling algorithm. Section III shows the simulation results. Finally, we conclude our work in Section IV.

II. THE PROPOSED APPROACH

A. System Architecture

In this paper, a scheduling-based resource reservation and flow scheduling algorithm is proposed. The proposed method allows end-to-end resource reservation for cloud UE, which means resources are reserved in all of the intermediate nodes between the sender and the receiver UE. 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's flexible. **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. Scheduling period of each forwarding device and Access Point (AP) is classified 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, normal best-effort UE's flows will be served.

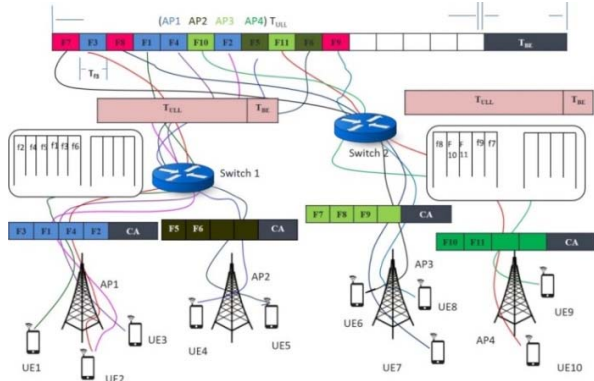


Fig. 1 System Architecture

B. Resource reservation

The proposed MAC protocol is composed of two different access methods: a Contention Free Access (CFA) and Contention Access (CA). CFA is intended for transmissions of delay sensitive traffics and CA is intended for transmissions of best-effort traffics. The CFA method is also used to send resource reservation requests (RRR). The CFA is divided into periodic Time division multiple access (TDMA) frames, each consisting of F time-slots, and each time-slot supports a fixed maximum throughput per time-slot. The legacy CSMA/CA is used to access the channel in CA method. It requires the UE to perform time synchronization with the central SDN controller before sending RRR messages. The controller maintains this synchronization for the UE across all APs regardless of the UE movement. The reason behind is to eliminate the complex time synchronization process performed by the UE every time it moves to different AP.

The frame architecture of proposed MAC protocol is shown in **Fig. 2**. Time is divided into a series of fixed-length periods called frame, in each frame a CFA followed by a CA; CFA is further divided into fixed time-slots. As mentioned before, CA is only used for time-slots reservation and for transmissions of

BE traffics. And then the multiple data packets are transmitted during the CFA.

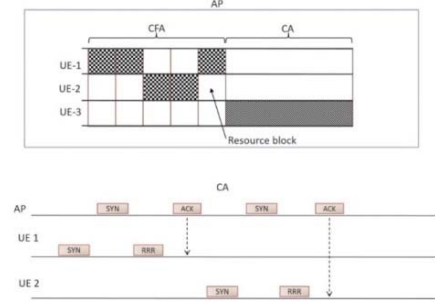


Fig. 2 The Resource reservation method

All UEs with traffic requirements reserve multiple periodic time-slots for its data packets transmissions with only one successful handshake in CA. In the proposed MAC protocol, there are two kinds of control packets, Resource Reservation Request (RRR) packet and Acknowledgement (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 adds 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. In addition, reservation table is maintained and updated at each UE through RRR-ACK negotiations, which includes the current time-slots usage condition.

C. Flow Scheduling

Consider a flow defined by a traffic specification with peak rate, the maximum packet size, the sustainable rate, and the maximum burst size, TSPEC (p, M, r, b). Network Calculus provides us the following definitions to calculate the effective bandwidth (eB) of a flow [6]. For a flow with T-SPEC (p, M, r, b), the effective bandwidth is:

$$eB = \text{MAX} \left\{ \frac{M}{D}, r, p \left(1 - \frac{D-M}{p-r} \right) \right\} \quad (1)$$

where D is the delay requirement of the flow.

Let the time axis be divided into scheduling frames, each consisting of F time-slots, where each time-slot supports the maximum throughput of $T_{\text{throughput}}$. Let $R(f)$ denote the number of time-slot reservations per scheduling frame needed to support the flow, where $0 \leq R(f) \leq F$. We can calculate $R(f)$ using the following equation.

$$R(f) = \frac{eB(f)}{\text{floor}(\frac{T_{\text{throughput}}}{M}) * M} * F \quad (2)$$

The Generalized Processor Sharing (GPS) concept presents a theory for computing packet departure times in [5]. 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 $\tau^*(f)$ denote the ideal inter-departure time

(IIDT) between adjacent packets in a perfectly scheduled flow with zero jitter.

$$\tau^*(f) = \frac{F}{eB(f)} \quad (3)$$

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

$$D^*(i, f) = D^*(i - 1, f) + \tau^*(f) \quad \text{for } i > 1 \quad (4)$$

$$D^*(i, f) = \text{Rand}(1, \tau^*(f)) \quad \text{for } i = 1 \quad (5)$$

Where, $\text{rand}(\cdot)$ is a random function for uniform distribution.

The proposed algorithm introduces the Virtual Finishing Time (VFT) of each packet in a flow to give priority to delay sensitive flows. Hence, we can use the equation (4) and (5) to calculate the VFT of each packet for a given flow. The proposed flow scheduling algorithm starts its works by first 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 will be updated by adding IIDT value of that flow to its current VFT value. **Fig. 3** shows the proposed flow scheduling algorithm.

Algorithm: Flow Scheduling

```

For each flow  $i$  in the controller
  Calculate  $R(f)$ 
  Calculate  $\tau^*(f)$ 
  Calculate the initial VFT =  $\text{Rand}(1, \tau^*(f))$ 
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. 3 The proposed flow scheduling algorithm

III. EXPERIMENT RESULTS

In this section, we present the numerical results and the performance evaluation. Five UEs associated with a single AP are considered and each UE will generate one flow with different TSPEC parameters. The Generic Cell Rate Algorithm (GCRA) is used to shape the traffic of each flow according to the TSPEC parameters [8]. To analyze the performance of the proposed scheduling algorithm, different system bandwidth of APs is considered. To be specific, 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 (The proposed, WFQ and round-robin) in the scenario, in which the system bandwidth is limited, enough and surplus.

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	50 K
Throughput per Time-slot	20Kb, 30Kb and 40Kb

A. Scenario configuration

Safety-critical traffic always is characterized by very-low frequency of occurrence, extremely short delay constraints, and small packet size. The proposed algorithm satisfies the requirement of such traffics best compared to the other two. In equation (1), it's 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 option of equation (1). Similarly, the second configuration contains five flows but, the eB is equal to the second option of equation (1). In another word, the two scenario *configuration* above 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	P (Kbps)	M (Kb)	r (Kbps)	b (Kb)	Delay (Sec)	W	eB (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	P (Mbps)	M (Kb)	r (Kbps)	B (Kb)	Delay (Sec)	W	eB (Kbps)
1	6	12	60	80	0.5	3	60
2	3.4	2	90	152	0.1	4	90
3	1.26	8	900	37	0.5	6	900
4	5.64	10	600	100	0.2	5	600
5	2.16	11	90	60	0.3	4	90

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

B. Simulation results

Fig. 4 shows the average throughput of all five flows in the first and second scenario configuration. The figure on the left illustrates the result of each 3 scheduling algorithms in terms of average throughput. As you can see from the **Fig. 4**, 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. The figure on the right illustrates the result for the second scenario. The proposed algorithm has better average throughput for flow 3 and flow 4 (they have a 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.

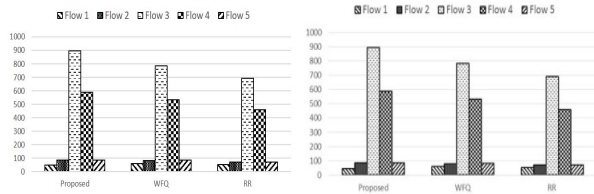


Fig. 4 The average throughput of first scenario configuration

Fig. 5 shows the time-slot utilization of each three scheduling algorithms. The figure on the left indicates that the proposed algorithm has better time-slot utilization for AP with 2Mbps bandwidth. The figure on the right indicates that the proposed algorithm has better time-slot utilization for AP with 2Mbps bandwidth for the second scenario configuration. Keep in mind that the maximum achievable throughput of this configuration is 40Kb per time-slot.

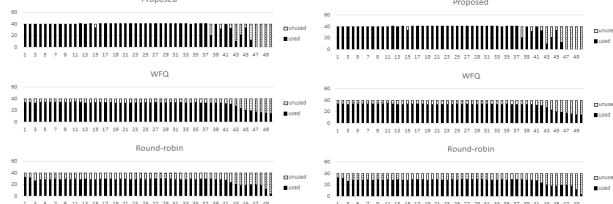


Fig. 5 The time-slot utilization

Fig. 6 shows the system throughput comparison between the three AP capacities for the two scenario configurations. This figure illustrates that the proposed scheduling algorithm has better system throughput in each AP capacities for both.

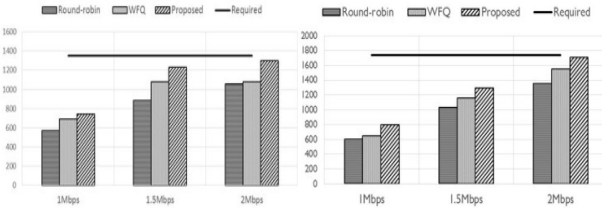


Fig. 6 System throughput comparison

From the above 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.

IV. CONCLUSION

In this paper, we proposed a scheduling-based resource reservation and flow scheduling algorithm for cloud UE. 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. This method does not consider solely the maximum packet size or the rate of the flow, but also it makes decision based on the effective bandwidth of the flow. Simulation results show it has capabilities to satisfy the QoS requirements of delay sensitive flows and improve the average throughput for delay-sensitive flows.

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] A. Gupta and R. K. Jha, "A survey of 5G network: architecture and emerging technologies," *IEEE Access*, vol. 3, no. 1, 2015, pp. 1206-1232.
- [2] J. G. Andrews et al., "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, June 2014, pp. 1065-1082.
- [3] Z. Meng, W. Chanle and Y. Gang, "Dynamic advance reservation for network resource," *Third International Conference on Intelligent System Design and Engineering Applications*, Hong Kong, 2013, pp. 618-621.
- [4] G. Boudour, C. Teyssié and Z. Mammeri, "Scheduling-based reservation MAC protocol for bandwidth and delay optimization in wireless mesh networks," in *Proceedings of IEEE International Conf. on Wireless and Mobile Computing, Networking and Comm.*, Avignon, 2008, pp. 272-277.
- [5] 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, Apr 1994, pp. 137-150.
- [6] F. Hu, Q. Hao and K. Bao, "A survey on software-defined network and open-flow: from concept to implementation," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, 2014, pp. 2181-2206.
- [7] 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.
- [8] Perros, H.G., *An introduction to ATM networks*. 2002: John Wiley & Sons.