Scalable and Adaptive Resource Scheduling in IEEE 802.16 WiMAX Networks

Hanwu Wang
Department of Computer Science,
City University of Hong Kong,
Hong Kong, SAR, China
whanwu2@student.cityu.edu.hk

Weijia Jia
Department of Computer Science,
City University of Hong Kong,
Hong Kong, SAR, China
wei.jia@cityu.edu.hk

***Abstract*—This paper proposes the so-called adaptive resource scheduling (ARS) schemes to cope with both uplink and downlink traffic transmissions in WiMAX, in order to fully utilize the radio resource, and meet the various QoS requirements as well. Specifically, for the uplink transmission, each SS control and adjust its local sessions in a distributed manner; while for downlink the BS performs the resource scheduling in a centralized manner. The corresponding rate control mechanism is given, by adopting the control theory method as well. The simulation results also validate the efficiency and practice of our proposed scheme.**

Keywords-WiMAX; rate contro; stability

I. INTRODUCTION

The worldwide interoperability for microwave access (WiMAX) [1-4], a new technology and solution for broadband wireless access networks, is developed by the IEEE 802.16 standard working group with rapid growth over past several years. WiMAX is a promising alternative to provide last-mile access in wireless metropolitan area network (WMAN) with the advantages of high speed, low cost, rapid and easy deployment. So that a large number of applications can be supported even in the areas where the installation of wired infrastructure is economically or technically infeasible. A WiMAX network consists of base station (BS) and subscribe station (SS). The former has wired connection with external network (e.g., Internet) and acts as a gateway for its internal SSs, while the later is served as access point to aggregate various applications from end users in its local area. Although the physical (PHY) layer and medium access control (MAC) signaling mechanisms have been well defined in the 802.16 standard specifications, the radio resource scheduling and management, which are regarded as the crucial components to achieve the desired QoS performance requirements, still remain as open issues. In [5] a dynamic bandwidth allocation scheme is proposed for broadband access wireless networks, however, the QoS differentiation was not addressed. In [6] a priority-based scheduling mechanism by employing AMC (adaptive modulation and coding), is proposed to meet with various QoS requirements, however only a single user scenario is considered. Both Niyato [7] and Rong [8] address the resource scheduling and admission control in detail and construct the relevant optimal models aiming at maximizing system utility. Sayenko [9] presents scheme to allocate resource (slots) to various service types in a certain order, but call admission control is not considered.

In this paper, we study the radio resource scheduling and traffic management in WiMAX networks from a new angle, specifically, we first distinct the uplink transmission from the downlink transmission for the whole network sessions. Then for the uplink scheduling, we let the different subscribe stations (SS) adaptively deal with its local session in a distributed manner, also with the coordination of base station (BS); For the downlink scheduling, the BS deals with the resource allocation and transmission control in a centralized manner. Our proposed scheme is scalable in the sense that different types of sessions can be integrated into a unified scheduling process to satisfy a flexible QoS (Quality of Service) requirement. Our scheme is adaptive that all the traffic rates can be controlled and tuned fast under different network traffic-load conditions. We adopt the classical control theory method into our proposed mechanism, and to the best of our knowledge it is the first time to apply such method in WiMAX system, which helps to achieve high efficiency (utilization), perfect traffic throughput, fairness, and system stability.

There are five types of the traffic service defined in WiMAX system, namely UGS (*Unsolicited Grant Service*), rtPS (*Real Time Polling Service*), ertPS (*Extended Real Time Polling Service*), nrtPS (*Non- Real Time Polling Service*), and BE (*Best Effort*) respectively. The UGS rtPS and ertPS provide a guaranteed service without contention. The nrtPS, though holding a periodical polling by BS, also need to compete with BE (most widely used service). Therefore this paper aims to schedule all the competing traffics using the available resources, in order to meet the desired performance requirements.

II. OUR SOLUTION: ADAPTIVE RESOURCE SCHEDULING (ARS)

In this section, we present our scheme, named adaptive resource scheduling (ARS), to deal with all the data traffics in WiMAX networks, and the relevant traffic scheduling model is illustrated in Figure 1.

As shown in Figure 1, the data traffics in the WiMAX system can be divided into two parts, namely uplink (from SSs to BS) traffic and downlink (from BS to SS) traffic

978-1-4244-2324-8/08/$25.00 © 2008 IEEE.

respectively. For the uplink, we concern how the traffics from each SS are scheduled, regardless of where they will destine (maybe to Internet or to the other SSs within the same cellular). In the downlink, we concern how the traffics, which have been aggregated at the BS, are transmitted out of it, without the consideration of where they are from (maybe from the Internet or the other SSs). Assume all the SSs (including BS) have an infinite transmitting or receiving capacity. Therefore the key performance requirement is how to allocate bandwidth for all the greedy SSs for both transmissions. Correspondingly our proposed ARS can be formulated as two independent processes for uplink and downlink respectively.

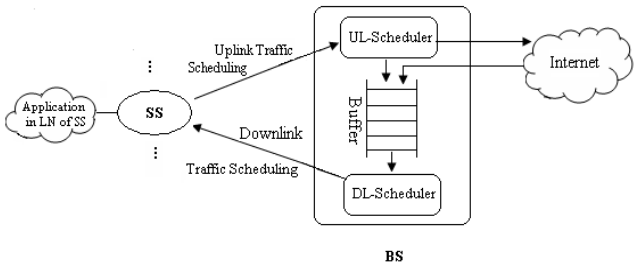

Figure 1 The traffic scheduling model under ARS

As for the uplink scheduling, each SS will control and adjust traffic rates for its local sessions (applications) in a distributed manner; and the control module UL-Scheduler in BS, is to decide the uplink bandwidth assignment for SSs in a periodical manner. Specifically, the UL_Scheduler in BS should first reserve bandwidth for those guaranteed services (e.g. UGS), then it fairly allocate the available bandwidth among all the competing SSs. Note that each SS will periodically report its traffic load (application type, quantity etc) to UL_Scheduler by sending the request messages (denoted by *Request*$_i$ as in Figure 2), to invoke the BA (bandwidth allocation) module and gets its fair portion of the available bandwidth according to

$$B_u(i) = (w_i \times B_u) / \sum w_i \tag{1}$$

Where B_u is the total available uplink bandwidth, w_i is the weight of SS_i, which is determined by BS according to its reported local traffic load; and $B_u(i)$ is the available bandwidth assigned to SS_i. Note that the bandwidth assignment is broadcasted to each SS by BS using UL_MAP messages in a periodical manner. Based on it, any of the SSs, say SS_i, can schedule its local m traffic sessions ($i1$, $i2$, … im) passing through it in a distributed manner, and these sessions will be integrated into a flow i and further transmitted upwards to BS using the assigned uplink bandwidth $B_u(i)$. Suppose we take an infinite small sampling time period, say T, then the dynamic of buffer occupancy between two consecutive T intervals at SS_i can be described by (2)

$$Q_i[(n+1)T] = [Q_i(nT) + \sum_{j=1}^{m} s(ij) T\lambda_{ij}(nT - \tau_{ij}^f T) - B_u(i)]^+ \tag{2}$$

Where $[x]^+ = \max(0, x)$, $T\lambda_{ij}(nT)$ is date traffic sent to SS_i by session ij (with the traffic rate $\lambda_{ij}(nT)$)during the nth interval of T, $\tau_{ij}^f T$ is the forward delay from this session source to SS_i, where τ_{ij}^f is an integer. $s(ij)$ is the traffic flag indicating whether this traffic is active (equal to one) or not (equal to zero). $Q_i(nT)$ denotes buffer size at time nT. After we remove the non-linear constrain and use the abbreviate notation, (2) can be rewritten by (3)

$$Q_i(n+1) = Q_i(n) + \sum_{j=1}^{m} s(ij)\lambda_{ij}(n - \tau_{ij}^f) - B_u(i) \tag{3}$$

Where $Q_i(n)$ and $\lambda_{ij}(n)$ stand for $Q_i(nT)$ and $\lambda_{ij}(nT)$ respectively. As for the rate control for all the competing sessions at SS_i, we design the following proportional controller

$$R_i(n) = B_u(i) - k(Q_i(n) - Q_{i0}) \tag{4}$$

Note $B_u(i)$ is the available bandwidth (maximal traffic rate) for all the competing sessions passing through SS_i, and $Q_i(n)$ is the buffer occupancy at SS_i as described above, Q_{i0} is the desired threshold of the buffer size, k is the control parameter. $R_i(n)$ is total allowed traffic rate for all these sessions.

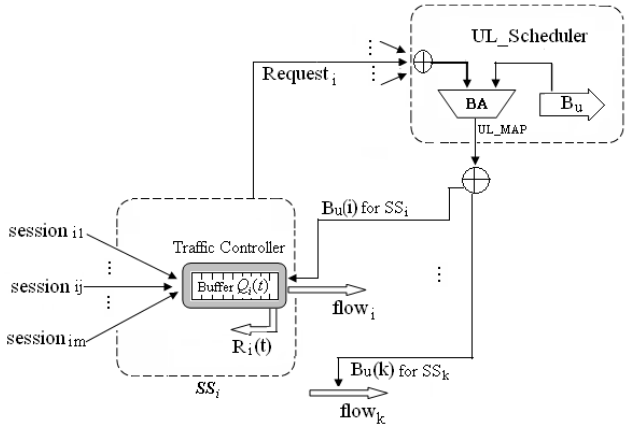

Figure 2 The uplink traffic management at SS_i as well coordinated by BS

On the other hand, as for the downlink traffic scheduling, the BS is the only transmitter and this work is quite different from the uplink. Specifically the BS takes charge of all the downlink traffic scheduling in a centralized manner. We design a corresponding DL-Scheduler at BS in order to deal with the downlink traffics passing through it. Specifically, we classify the incoming traffics for downlink transmission according to their service type, and the relevant buffers namely bu_G, bu_n and bu_b, are set for aggregating the Guaranteed, nrtPS and BE traffics respectively. While $r_G(t)$, $r_n(t)$ and $r_b(t)$ are the aggregated traffic rate out of each bu_i (i=G, n, b), which are determined by the DL_Scheduler (as in Figure 3). Similarly, we can formulize the buffer dynamics $Q(t)$ at BS by

$$Q(n+1) = Q(n) + r_G(n - \tau_G^f) + r_n(n - \tau_n^f) + r_b(n - \tau_b^f) - B_d \tag{5}$$

Where τ_G^f, τ_n^f and τ_b^f are the forward delay from bu_i to DL-Scheduler respectively. Moreover, we can also design a rate controller as in the uplink case to determine the traffic rate $R(t)$ for all the downlink competing sessions. Note that the aggregated traffic flow out of each bu_i (i=G, n, b) will be

divided into different SS sessions, hence the Traffic Controller (see Fig. 3) not only control the aggregated rate $r_i(t)$ (i=G, n, b), but also determine the inner allocation of each aggregated traffic for different SSs. As for the bandwidth (rate) allocation policy, the guaranteed traffics do not participate in the bandwidth contention. While the nrtPS and BE traffic rates are controlled and tuned according to their traffic load and the availability of the bandwidth.

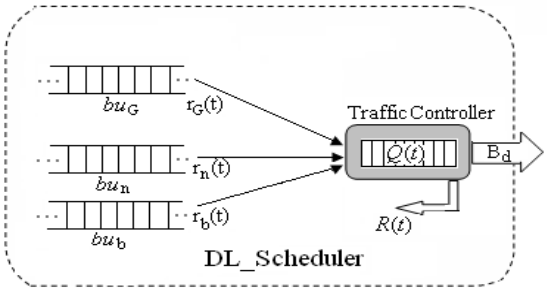

Figure 3 The downlink traffic management at BS

III. RATE CONTROL AND ADJUSTMENT

A. Rate Control Algorithm

The previous section presents the basic framework for traffic scheduling. In this section, we further give the detail mechanism for rate allocation and control to meet the desired performance requirements.

As for the uplink transmission, we perform the rate allocation and control among all the competing sessions at SS_i according to Algorithm 1.

Algorithm1 Rate Control for uplink traffic (at SS_i)
At the beginning of each frame interval T_f
$B_A(t)=B_u(t)-\sum s(ij)*MIB(ij)$ *//allocate MIB(s)*
If($B_A(t)<0$)
Select ($Min\{rw(ik)\}$) ; // *block some nrtPS session(s)*
$rw(ik) \leftarrow rw(ik)+1$; *//update its(their) rw(s)*
$s(ik) \leftarrow 0$; *//reset its(their) active flag(s)*
If ($B_A(t)=0$) *//no bandwidth for competing*
Feedback active MIBs to their sources; *Else*
$R_i(t) \leftarrow B_A(t)-k(Q_i(t)-Q_{i0})$; *//P-Controller mechanism*
For each active competing session
$R_{ij}(t) \leftarrow w_{ij} * R_i(t)/ w_i$; // *fair share for each one*
For the nrtPS sessions
$R_{ij}(t) \leftarrow R_{ij}(t) + MIB(ij)$;
Feedback $R_{ij}(t)$ *to the source of each session*;
At the end of each frame interval T_f
Update rw value of the just served sessions;
Select ($Max\{rw_{ik}\}$);
$s(ik) \leftarrow 1$; *//reactive some blocked nrtPS sessions*
End

Specifically, Algorithm 1 is periodically performed on a WiMAX frame duration basis, which is denoted by T_f. Note T_f is different from (greater than) the sample time T for buffer dynamics as described above. $B_u(t)$ is the uplink bandwidth assigned to SS_i during each frame, $MIB(ij)$ is minimal bandwidth requirement for the nrtPS traffic ij, so $B_A(t)$ is the residual bandwidth for all the competing sessions, and $R_i(t)$ is allowed traffic rates based on $B_A(t)$ according to the control theory mechanism (4). Note that it's quite often that bandwidth is not sufficient for all the traffic transmission at SS_i, and the nrtPS traffics always take higher priority over the BE traffics. Hence the BE traffics have to be blocked and wait for the available bandwidth for their transmissions. Moreover, not all the nrtPS sessions can simultaneously be satisfied when facing the heavy congestion. In this respect, we may let all the nrtPS sessions be scheduled in turn when the system is overloaded, so that each one can get the chance to be served on a regular basis to meet nrtPS QoS requirement. Specifically, we set a connection priority vector $\boldsymbol{rw}$ = (rw (i1), rw (i2), ..., rw(ik)) for all the competing nrtPS sessions at SS_i, and rw(ik) denotes the priority for an nrtPS session to be admitted by SS_i. As in Algorithm 1, at the beginning of every frame at SS_i, the nrtPS session(s) with the minimal rw are selected to be blocked due to the lack of resource, meanwhile their rw (priority) are increased by one for the next contention. At the end of every frame, all the just served session will decrease their rw by one, and the blocked sessions with the higher rw will be selected for bandwidth allocation. Under such way, each nrtPS session has the fair chance to be admitted by SS_i and get its minimal bandwidth requirements on a regular basis.

When there is available bandwidth after satisfying the MIB requirements of nrtPS sessions, the BE traffic sessions are invoked quickly and get their service at the best effort, then SS_i using the P-Controller method to further control and tune data rates for the competing BE and nrtPS traffics.

As for the downlink case, we first serve the guaranteed traffics in bu_G, and then satisfy the MIB requirements of the nrtPS session in bu_n as is done in the uplink. Based on these, the residual available bandwidth for competition is $B_A(t) = B_d - r_G(t) - \sum MIB(i)$, where $MIB(i)$ is for the nrtPS traffics in bu_n. We further determine the total traffic rate, i.e., $R(t) = B_A(t) - k(Q(t) - Q_0)$ for all competing sessions by employing the control theory method as in Algorithm 1. Specifically, $R(t)$ is divide into the $r_n(t)$ and $r_b(t)$ for these two aggregated traffics according to their weights, and the $r_i(t)$ (i=n,b) is shared by its inner sessions of different SSs.

B. Stability Analysis

For more specifically, we further make the stability anaysis about the P-Controller mechansim presented in our ARS. We take a subscribe station SS_i as representative for uplink analysis. This process can be easily extended to the downlink case.

Based on (4), the traffic rate for any competing traffic session ij at SS_i is given by

$$\lambda_{ij}(n) = w_{ij} \times [B_u(i) - k(Q_i(n - \tau_{ij}^b) - Q_{i0})] \quad (6)$$

Where τ_{ij}^b is the delay from SS_i back to this traffic source, and w_{ij} is the weight for it. Take the Z-transform of (3) and (6), we further have (7) and (8)

$$(Z-1)Q_i(Z)=\sum_{j=1}^{m}s(ij)\lambda_{ij}(Z)Z^{-\tau_{ij}^f}-\frac{ZB_u(i)}{Z-1} \quad (7)$$

Where $Q_i(Z)=\sum_{n=0}^{\infty}Q_i(n)Z^{-n}$; $\lambda_{ij}(Z)=\sum_{n=0}^{\infty}\lambda_{ij}(n)Z^{-n}$.

$$\lambda_{ij}(Z)=\frac{w_{ij}(B_u(i)+kQ_{i0})Z}{Z-1}-kw_{ij}Q_i(Z)Z^{-\tau_{ij}^b} \quad (8)$$

Substitute (8) into (7), one yields

$$(Z-1)Q_i(Z)=\sum_{j=1}^{m}s(ij)w_{ij}[\frac{(B_u(i)+kQ_{i0})Z^{1-\tau_{ij}^f}}{Z-1} - kQ_i(Z)Z^{-\tau_{ij}}]-\frac{ZB_u(i)}{Z-1} \quad (9)$$

Where $\tau_{ij}=\tau_{ij}^f+\tau_{ij}^b$, after the shift manipulation, one yields

$$[(Z-1)+k\sum_{j=1}^{m}s(ij)w_{ij}Z^{-\tau_{ij}}]Q_i(Z)= \sum_{j=1}^{m}\frac{(B_u(i)+kQ_{i0})Z^{1-\tau_{ij}^f}}{Z-1}-\frac{ZB_u(i)}{Z-1} \quad (10)$$

Taking the characteristic polynomial equation of (9), one yields the following

$$Z-1+k\sum_{j=1}^{m}s(ij)w_{ij}Z^{-\tau_{ij}}=0 \quad (11)$$

According to the control theoretic principle [10], when all the zeros of the above equation (11) lie within the unit disc by choosing a (set of) proper k, the closed loop system (3) and (6) is asymptotically stable, so that various traffic throughput, buffer queue length and delay at SS_i will reach and be kept at the desired levels. The stability analysis process for the P-Controller mechanism in the DL_Scheduler is similar to the uplink analysis.

IV. SIMULATIONS

In this section, we construct the relevant simulation models to valid our proposed ARS. The first configuration is to test the uplink traffic scheduling, while the second one is used for the downlink evaluation. We do not consider guaranteed services. Both models take a one-BS with three-*SS* topology. Each SS only communicates with the BS. The bandwidth for both uplink and downlink is set to 100Kbps, and the threshold of buffer size at each SS_i (i=1, 2, 3) and BS is set to 10 and 20 packets (one Kbits per packet) respectively. The transmission delay on each link set to 2ms.

As for the first model, the whole simulation time is set to 500ms and divided into several phases. Specifically, they are T_1(0, 100ms) where $SS_{1\text{-}3}$ each station serves a BE session; T_2(101, 200ms) where three nrtPS sessions with a *MIB* of 9Kbps join in SS_1, SS_2 and SS_3 respectively; T_3 (201, 300ms) where three BE traffics join in SS_1, SS_2 and SS_3 respectively; T_4 (301, 400ms) where three nrtPS sessions with a *MIB* of 12Kbps join in SS_1, SS_2 and SS_3 respectively; T_5(401, 500ms) where the first three nrtPS session in $SS_{1\text{-}3}$ go off. Note that we suppose each competing session has an equal weight, except for the *MIB*s of the nrtPS traffics. The simulation results, i.e., traffic rates for each session and buffer occupancy at each SS_i, are shown in Figure 4 and 5 respectively. Note that R_{ij} denotes the jth traffic rate (throughput) at SS_i, while Q_i denotes its buffer queue size.

As for the second simulation model, the whole simulation time is set to 600ms and divided into several phases. Specifically, they are T_1(0, 100ms) where two BE traffics and one nrtPS traffic with *MIB* of 20Kbps occur in bu_b and bu_n respectively; T_2(101, 200ms) where a new nrtPS session with *MIB* of 15 joins in bu_n; T_3(201, 300ms) where a new BE session occurs at bu_b; T_4(301, 400ms) where a new BE session starts at bu_b; T_5(401, 500ms) where the nrtPS session at bu_n departs; T_6(501, 600ms) where a new nrtPS session with *MIB* of 10Kbps and BE session join in the bu_n and bu_b respectively. Note that R_{ij} (i=n, b) denotes the jth traffic rate at bu_i, and Q denotes the buffer queue size at the BS.

As we can see from Figure 4 and 5, during T_1 SS_1, SS_2, SS_3 get an equal share of the available bandwidth for their BE traffics; the buffer occupancy at each SS is kept at the target value. While from T_2 to T_5 BE sessions always give way to *MIB* requirement of nrtPS sessions, and then compete for the residual available bandwidth. The traffic rate of each session at different stages is stabilized at the desired value, so does the buffer size (occupancy) at each SS.

As for the downlink case (see Figure 6 and 7), from T_1 to T_6, the *MIB* requirement of each nrtPS session in bu_n are satisfied first, after this the BE sessions in bu_b get chance to compete for the left available bandwidth. The traffic rates in each bu_i is stabilized near the desired value, and the buffer occupancy $Q(t)$ is also maintained at the target value.

V. CONCLUSION

This paper proposes a scalable and adaptive scheme namely ARS, to deal with both the uplink and downlink traffic transmissions in WiMAX system respectively, so as to achieve the high bandwidth utilization, good traffic throughput, and various desired performance requirements among different competing users. While the uplink scheduling is performed at each SS in a distributed manner, with the coordination of the BS; the downlink scheduling is executed at the BS in a centralized way. The relevant algorithm for bandwidth assignment and rate control is presented in detail; the stability analysis is also performed to get the desired control parameter. Simulation results also validate the efficiency and scalability of our proposed scheme.

ACKNOWLEDGMENT

The work is fully supported by RGC General Research Fund (CERG) no. 9041129 (CityU 113906), HK SAR and CityU Strategic Research Grant no. 7002214.

REFERENCES

[1] Fundamentals of WiMAX: understanding broadband wireless networking. Jeffrey G. Andrews, Arunabha Ghosh, Rias Muhamed. Prentice Hall, 2007.

[2] WiMAX: technology for broadband wireless access. Loutfi Nuaymi, Chichester, England; Hoboken, NJ: John Wiley, 2007.

[3] IEEE Std 802.16a-2003, "IEEE Standard for Local and metropolitan area networks--Part 16: Air Interface for Fixed Broadband Wireless Access Systems--Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz," 2003.

[4] IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," 2004.

[5] G. Li and H. Liu, "Dynamic Resource Allocation with Finite Buffer constraint in Broadband OFDMA Networks," Proc. IEEE Wireless Comm. And Networking Conf., vol. 2, pp. 1037-1042., Mar. 2004.

[6] Q. Liu, X. Wang, and G.B. Giannakis, "A Cross-Layer Scheduling Algorithm With QoS Support in Wireless Networks," IEEE Transactions on Vehicular Technology, vol. 55, no. 3, pp. 839-847, May, 2006.

[7] D. Niyato, E. Hossain, "A Queuing- Theoretic and Optimization- Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks", IEEE Transactions on Computers, vol. 55, no. 11, November 2006.

[8] B. Rong, Y. Qian, K. Lu, "Integrated Downlink Resource Management for Multiservice WiMAX Networks," IEEE Transactions on Mobile Computing, vol. 6, no. 6, June 2007.

[9] A. Sayenko, O. Alanen, J. Karhula, T. Hamalainen, "Ensuring the QoS requirements in 802.16 scheduling," Proceedings of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, Pages: 108 – 117, c2006

[10] Bubnicki, Zdzislaw**.** Modern Control Theory. Berlin ; New York : Springer, c2005

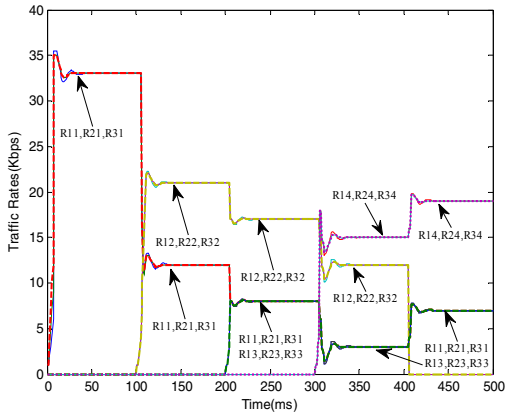

Figure 4 The traffic rate $R_{ij}(t)$ (i=1, 2, 3; j=1, 2, 3, 4) at each SS_i for the first model

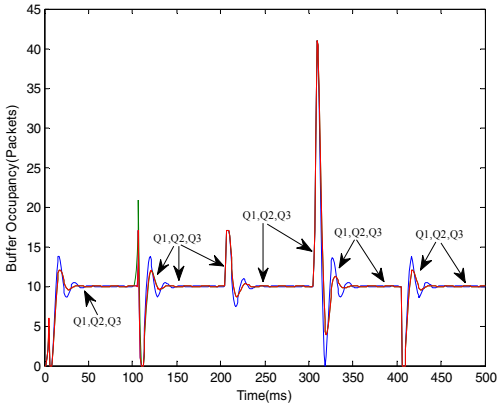

Figure 5 The buffer occupancy $Q_i(t)$ (i=1, 2, 3) at each SS_i for the first model

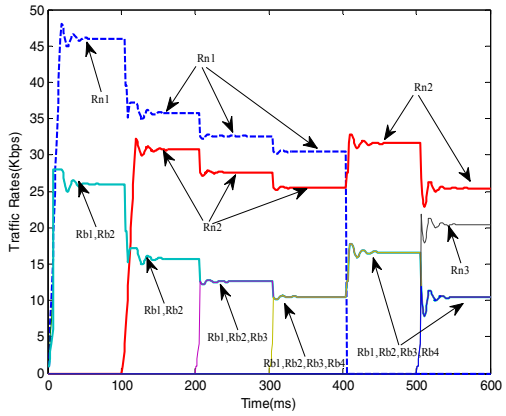

Figure 6 The traffic rate $R_{ij}(t)$ (i=n, b; j=1, 2, 3, 4) at each bu_i for the second model

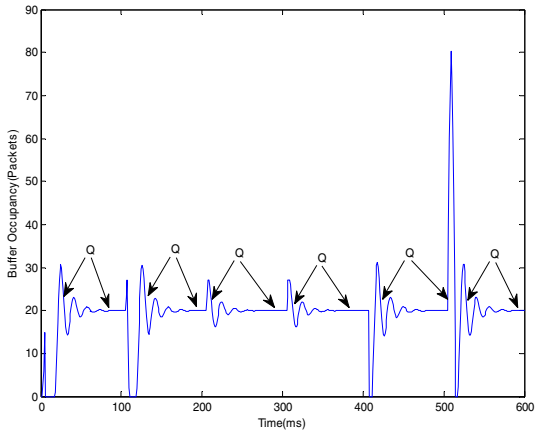

Figure 7 The buffer occupancy $Q(t)$ at the DL_Scheduler (BS) for the second model