QoS Scheduling for Heterogeneous Traffic in OFDMA-based Wireless Systems

Youngki Kim
Mobile R&D Laboratory
KT, Korea
Email: kyk@kt.com

Kyuho Son and Song Chong
School of EECS
KAIST, Korea
Email: {skio@netsys, song@ee}kaist.ac.kr

***Abstract*—In this paper, we propose a scheduling framework for heterogeneous traffic in OFDMA-based wireless systems. The proposed scheduling algorithm not only satisfies the QoS requirements of the real-time traffic but also maximizes the utility of the non real-time traffic. Step-by-step approach is used to achieve these two objectives simultaneously with low complexity and traffic class prioritization. A well-known bipartite matching algorithm and a standard gradient scheduling algorithm are adopted for the QoS scheduling of the real-time traffic and for the utility maximization scheduling of the non real-time traffic, respectively. Moreover, a noble *beta deadline parameter* is introduced to control the balance between the QoS provisioning and the diversity gain. Extensive simulation results in various scenarios are provided to demonstrate the good features of our scheduling framework.**

I. Introduction

Wireless systems are moving towards supporting the wide range of heterogeneous services with different traffic properties. These services are the mixture of real-time flows (e.g., voice/video and multimedia) and non real-time best effort flows (e.g., file downloads or web browsing). From a service provider's perspective, one would like to maximize revenue as many as possible by making customers satisfied within the limited budget. From customers' perspective, they require the system to ensure their quality of service (QoS) of real-time flows and/or maximize their throughput or utility (user satisfaction) of best effort flows.

At the same time, one of the key access technologies in current and next generation wireless systems is OFDMA (orthogonal frequency division multiplexing access). Most of wireless standards such as IEEE 802.11/16/20 have adopted OFDMA as an access technology because it has high degree of flexibility in allocating resources and strong characteristics against frequency selective fading.

Therefore, there exists an urgent need for developing a scheduling framework to meet the demands posed by the convergence of heterogeneous services especially on the area of radio resource management in OFDMA-based wireless systems. Furthermore, this scheduling framework whose operating point can be flexibly adjusted by reflecting customers' changing tendencies and service providers' strategic decisions is essential. Motivated by these demands, we set three goals of this paper as follows.

- Can we provide QoS guarantee to the real-time traffic in multi-carrier wireless systems?
- Can we realize utility maximization of the non real-time traffic while providing QoS guarantee to the real-time traffic?
- Can we balance between QoS guarantee and utility maximization in a simple and organized manner?

A. QoS Scheduling and Related Works

Packet scheduling plays an important role in QoS provisioning by providing mechanisms for the resource allocation and multiplexing at the packet level to ensure different types of applications meet their service requirements. Utility theory provides the reasonable methods to formulate the relations between user experience and various network performance matrices of the non real-time traffic quantitatively [1]. Providing utility maximization of the non real-time traffic must be kept behind providing QoS guarantee to the real-time traffic. However, just giving strict priority to the real-time traffic is not an intelligent solution in terms of efficient wireless resource utilization. Thus, we need a well organized scheduling framework which can balance between QoS guarantee and utility maximization.

Reliable delivery of the real-time traffic over wireless networks provide researchers a lot of challenges as the they require both throughput and delay guarantee. In [2] and [3], M-LWDF (modified largest weighted delay first) and EXP (exponential) give priority to the flow with high HOL (head of line) packet delay and good channel condition. By properly weighting the decision metric, they can improve delay performance compared to other schedulers such as Max. C/I and PF (proportional fair) [4]. In [5], opportunistic scheduling algorithm for delay sensitive traffic in OFDMA-based wireless networks is proposed, which gives some priority to the flow with regard to packet delay and packet loss ratio. The authors proposed a scheduling and source control algorithm of real-time traffic by explicitly controlling average queue length to a target value determined by delay requirement in [6].

This work was partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract. This research was also supported by the Ministry of Knowledge Economy, Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2009-C1090-0902-0037).

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

The remainder of this paper is organized as follows. In Section II, we describe our system model and problem formulation. In Section III, we propose QoS scheduling framework for heterogeneous traffic in wireless systems. In Section IV and V, we validate the proposed scheduling framework through various scenarios and conclude the paper, respectively.

II. Model Description and Problem Formulation

Consider a single-cell OFDMA wireless system as shown in Fig. 1. We denote by $\mathcal{S}$ the set of all sub-channels in the system, $\mathcal{N}_{RT}$ and $\mathcal{N}_{NRT}$, the set of all real-time (RT) and non real-time (NRT) flows (interchangeably used with 'traffic' or 'users') in the system, respectively. We consider only downlink transmissions in the time-slotted system indexed by $t=0,1,\ldots$. The real-time flow, VoIP or MPEG, has its own QoS parameters such as maximum latency (or deadline) of individual packet and average traffic rate whereas non real-time flow has no explicit QoS parameters. In the system, at each slot, the proposed scheduler determines the sub-channel assignment based on each flow's current *channel quality*, *minimum average throughput* and *individual packet deadline*.

In this paper, we aim at proposing a scheduling framework that maximizes the weighted sum rate of non real-time flows while maintaining QoS constraints of real-time flows in each time slot with the equal power allocation assumption[1], i.e., solves the following optimization problem **P**:

$$\textbf{P:} \quad \max \sum_{i\in\mathcal{N}_{NRT}} \sum_{j\in\mathcal{S}} U_i'(\bar{r}_i(t))\mu_{ij}(t) \qquad (1)$$

$$\text{subject to} \quad \theta_i(t) \geq \pi_i(t), \quad \forall i \in \mathcal{N}_{RT}, \qquad (2)$$

where the derivative of utility function of flow i, $U_i'(\cdot)$, is used as a weight and $\mu_{ij}(t)$ is the achievable channel capacity when sub-channel j is assigned to flow i at time slot t; And $\bar{r}_i(t) = \frac{1}{t}\sum_{\tau=1}^{t}\sum_{j\in\mathcal{S}}\delta_{ij}(\tau)\mu_{ij}(\tau)$ is the long-term throughput for flow i up to time slot t, where $\delta_{ij}(\tau)$ is the 0-1 indicator of allocating the sub-channel j to the flow i or not. Since each sub-channel is occupied by only one user, we have the following OFDMA constraint, $\sum_{i\in\mathcal{N}_{RT}\cup\mathcal{N}_{NRT}}\delta_{ij}(t) \leq 1, \forall j \in \mathcal{S}$.

Another variables for real-time flows, $\theta_i(t)$ and $\pi_i(t)$, need to be explained specifically. First of all, $\theta_i(t)$ is the actual amount of data allocated to real-time flow i at time slot t and $\pi_i(t)$ is given by

$$\pi_i(t) = \min\left[\max\left\{M_i, R_i^{min}(t)\right\}, R_i^{max}(t)\right], \qquad (3)$$

where M_i is the minimum required average traffic rate of real-time flow i. And $R_i^{min}(t)$ is the required data rate to provision QoS guarantee and $R_i^{max}(t)$ is the maximum possible data rate of real-time flow i at time slot t, which can send all of the packets in the queue in one time slot. The following subsection describes the detailed procedure for choosing the the value of $R_i^{min}(t)$ properly based on the newly introduced *beta deadline parameter*.

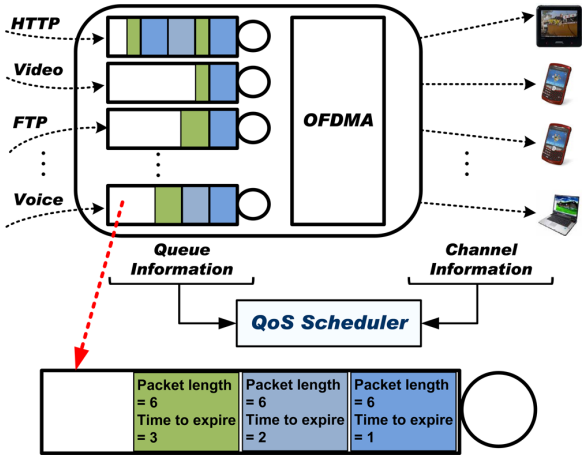


Fig. 1. System model

The problem **P** is not simple due to the complexity of mixing the non real-time flow's objective and the real-time flow's constraints. Therefore, we take a step-by-step approach to solve this problem: real-time QoS scheduling (in Section III-A) and non real-time scheduling (in Section III-B). We first allocate sub-channels to meet the QoS requirement of real-time flows, and then remaining sub-channels will be used to maximize the objective function of the non real-time traffic.

A. Beta Deadline Parameter

To explain *beta deadline parameter* we design, the snapshot of queue in the bottom of Fig. 1 is introduced. There exists three packets in this queue, where the length of each packet is 6 and time to expire values are 1 (unit time slot), 2 and 3, respectively. The required data rate, $\pi_i(t)$, for the QoS provisioning of the real-time traffic depends on how strictly the scheduling policy is imposed on the scheduler.

If we take an urgent scheduling policy[2] [9] which only considers the most urgent packets as scheduling candidates, then required data rate (per slot) is 6. If we take strict priority scheduling[3] [10] to provide higher priority to the real-time traffic than non real-time traffic, then the data rate should reach up to the value of sending all of the packets in the queue, that is, 18. We may take a policy somewhere between these two extreme cases. In [11], the authors set the required data rate as the sum of the packet length divided by each packet's time to expire value so that it can guarantee the maximum latency of each packet in average sense.

By looking into the process of calculating these required data rates carefully, we can draw an insightful connection between those policies just using one parameter β, which we refer as *beta deadline parameter*, where l_{ik} is the length of

[1]The equal power assumption has been frequently used for implementation simplicity as well as analytical tractability in downlink resource allocation problems [7]. Moreover, equal power allocation is near optimal in many cases especially in high SINR regime [8]

[2]Only the most urgent packets, whose timeout value is equal to one, will be scheduled for transmission in current time slot.

[3]The scheduler serves packets from priority level p only if there exists no packet of the queue of higher priority level than p.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

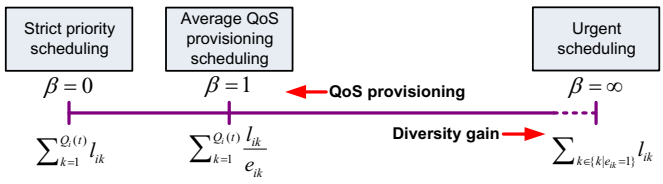


Fig. 2. Scheduling policy based on beta deadline parameter

the k-th packet of flow i, e_{ik} is the time to expire value of the k-th packet of flow i and Q_i is the total number of packets of real-time flow i at time slot t.

$$R_i^{min}(t) = \sum_{k=1}^{Q_i(t)} \frac{l_{ik}}{e_{ik}^{\beta}}, \tag{4}$$

Using this simple but well designed bridge, it is possible to convert the real-time traffic's QoS parameter (maximum latency) from time domain to the data rate domain, which makes the problem easy to solve.

Remark 1: Note that by setting β to ∞, 0 and 1, we can obtain previous three scheduling policies: the urgent scheduling policy [9], the strict priority scheduling policy [10] and the average QoS provisioning scheduling policy [11], respectively. Moreover, we can obtain arbitrary scheduling policies in between those above by choosing different values of $\beta \in [0, \infty)$.

Remark 2: We can balance the tradeoff between the QoS provisioning for the real-time traffic and diversity gain for the non real-time traffic with this *beta deadline parameter* as shown in Fig. 2. The higher (or lower) QoS provisioning for the real-time traffic is possible with the lower (or higher) value of β. However, the less (or more) leftover sub-channels lead to the less (or more) diversity gain for the non real-time traffic.

Remark 3: One could have questions on the rate conversion process in (4), because this framework inevitably causes the overhead of handling fragmented packets. By taking the deficit concept from [12], which was originally designed to give fairness to the flows which have variable size packets in wired network, we can effectively avoid packet fragmentation problem. We omit the details due to space limitation.

III. Proposed QoS Scheduling Framework

This section is devoted to describing our scheduling framework that consists of real-time QoS scheduling and non real-time scheduling. In the first step, we minimize the total number of allocated sub-channels to the real-time traffic while providing QoS guarantee. In the second step, remaining sub-channels are used to maximize the objective function of the non real-time traffic.

A. Step I: Real-time QoS Scheduling

In [11], the flow priority based max throughput channel selection algorithm decides the sub-channel assignment in decreasing order of each flow's priority and sub-channels are assigned to the selected flow until its QoS requirements are satisfied. However, because the higher priority flow can take preemptive selection of the channel which can give better channel capacity to the lower priority flow, this approach cannot fully utilize the channel diversity gain. In this paper, we consider a bipartite matching algorithm to fully exploit channel gain. Matching algorithm can be a efficient solution for sub-channel assignment problem when the required number of sub-channel is decided and the channel gain is known as discussed in On Kuhn's Hungarian method [13].

Initially, the number of sub-channels to be assigned to flow i at time slot t, $n_i(t)$, is obtained from the value of $\pi_i(t)$ divided by average $\mu_i(t)$. And then, we can formulate the following maximum weighted bipartite matching (MWBM) problem to find the sub-channel allocation matrix, where $n_i(t) = \left\lceil \frac{\pi_i(t)}{\bar{\mu}_i(t)} \right\rceil$, the number of sub-channels to be assigned to flow i at time slot t, and $\bar{\mu}_i(t) = \frac{1}{|\mathcal{S}|} \sum_{j\in\mathcal{S}} \mu_{ij}(t)$, the average sub-channel capacity of the flow i if all the sub-channels are allocated to it.

$$\max \quad \sum_{i\in\mathcal{N}_{RT}} \sum_{j\in\mathcal{S}} \delta_{ij}(t)\mu_{ij}(t) \tag{5}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{S}} \delta_{ij}(t) = n_i(t), \tag{6}$$

$$\sum_{i\in\mathcal{N}_{RT}} \delta_{ij}(t) \le 1. \tag{7}$$

Using the Hungarian method, we can find the sub-channel allocation matrix that can maximize the sum throughput under the constraint of providing the required number of sub-channels to each flow i. However, the result of maximum weight bipartite matching algorithm using average sub-channel capacity cannot give exact number of sub-channels to the flows. The following describes a detailed procedure to adjust the gap between the initial result of matching algorithm and the exact required number of sub-channels.

1: **Initialize**:
– (Over requested sub-channel handling) If required sub-channels are more than total number of sub-channels, normalize the required sub-channels:
if $\sum_{i\in\mathcal{N}_{RT}} n_i(t) > S$, then $n_i(t) = S \cdot \left\lceil \frac{n_i(t)}{\sum_{i\in\mathcal{N}_{RT}} n_i(t)} \right\rceil$.
– Solve the MWBM problem according to (5) ∼ (7).
2: **Repeat**: Adjusting the assigned sub-channels.
– If QoS is over provisioned, remove sub-channels from the over provisioned real-time flow.
– If QoS is under provisioned, add sub-channels to the under provisioned real-time flow.
3: **Finish**: Checking the termination condition.
– If there is no flow which is over allocated or under allocated sub-channels, then stop.
– Else if there are no more sub-channels available, then stop.
– Else, go to repeat step.

B. Step II: Non Real-time Scheduling

After the real-time QoS scheduling, the remaining sub-channels ($\mathcal{S}'$) will be used for the utility maximization schedul-

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

ing of the non real-time traffic in each time slot.

$$i^* = \arg \max_{i \in \mathcal{N}_{NRT}} U'_i(\bar{r}_i(t))\mu_{ij}(t), \quad \forall j \in S', \quad (8)$$

$$\text{where } U_i(\bar{r}_i(t)) = \log(\bar{r}_i(t)). \quad (9)$$

Various utility functions can be used according to the scheduling policy for the non real-time traffic. One of the well defined utility function is based on alpha proportional fairness proposed in [14].

$$U^\alpha(x) = \begin{cases} (1-\alpha)^{-1}x^{1-\alpha}, & \text{if } \alpha \neq 1 \\ \log(x), & \text{otherwise.} \end{cases} \quad (10)$$

In (8), general utility function is defined for $\alpha \geq 0$. In particular, $\alpha = 0$ is for maximum throughput, $\alpha = 1$ is for proportional fairness, and max-min fairness objective is achieved as α goes to ∞. Therefore, the α can be interpreted as a parameter of tradeoff between efficiency and fairness.

In this paper, we assume the utility function as $\log(\cdot)$ which allocates resources to each flows in a proportional fair manner. Using the $\log(\cdot)$ utility function in (8,9) for all the remaining sub-channels, the proportional fair scheduling for the non real-time traffic while exploiting multi-user and multi-channel frequency diversity is possible. It is worthwhile mentioning that the scheduling policy for the non real-time traffic can be replaced by any other well designed policies.

C. Desirable Features of Proposed Scheduling Algorithm

In this section, we emphasize good features of our proposed scheduling algorithm with reference to the desirable scheduling characteristics recommended in [15]. Our proposed scheduling algorithm has the following advantages.

Efficient link utilization: Efficient sub-channel resource utilization is achieved by adopting maximum weighted bipartite matching algorithm in QoS scheduling step and by considering sub-channel capacity in utility maximization scheduling step to exploit multiuser diversity.

Delay bound: Guaranteeing delay bound is achieved by studying the relationship between the individual packet deadline and the required data rate, and by assigning the necessary sub-channels to the real-time traffic in each time slot.

Fairness: Providing reasonable fairness to the non real-time traffic is achieved by designing multi-channel extension of utility maximization scheduling algorithm which realizes the proportional fair resource allocation.

Throughput: Average minimum throughput is guaranteed for the real-time traffic by considering this QoS parameter explicitly when deciding required data rate for the real-time traffic in each time slot.

Implementation complexity: Low implementation complexity is achieved by step-by-step approach which considers QoS scheduling for the real-time traffic first, and utility maximization scheduling for the non real-time traffic with the leftover sub-channels.

Graceful service degradation: Sub-channel over request handling algorithm is designed to degrade service quality gracefully by normalizing the amount of sub-channels when necessary sub-channels are larger than the total number of sub-channels due to bad channel quality or burst traffic input.

Delay/bandwidth decoupling: Beta deadline parameter is developed to control the balance between QoS guarantee and diversity gain. By setting different *beta deadline parameter* values according to the application types, we can control the relationship between delay and bandwidth strategically.

Flexible scheduling framework: Note that we construct the well organized scheduling framework for both real-time scheduling and non real-time scheduling with simple parameters for both traffic types. *Alpha proportional parameter* can balance between the tradeoff between efficiency and fairness for the non real-time traffic, whereas *beta deadline parameter* can balance between the QoS provisioning and the diversity gain.

TABLE I
SUMMARY OF SIMULATION PARAMETERS

Parameter	Value
Radius of cell	1km
User distribution	Uniform
Pass loss model	L=128.1 + 37.6$\log_{10}$R
Base station TX power	20W
System bandwidth	2.5Mhz
Number of sub-channels	24
Slot duration	2ms
Traffic type	Real-time (Voice, Video), BE

IV. Simulation Results

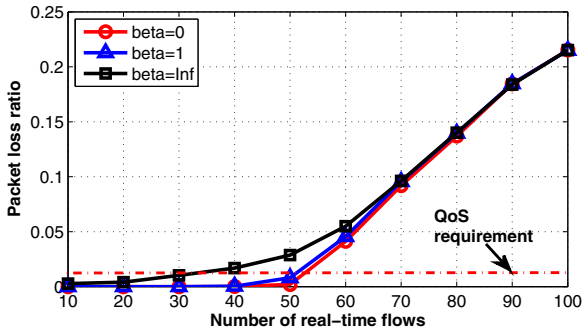
A. Simulation Environment

In this section, we will evaluate the performance of the proposed QoS scheduling framework to validate its characteristics as designed. We summarize the general simulation settings in TABLE I and the traffic models are presented in TABLE II, III. The frequency selective fading is generated by adapting ITU pedestrian B parameters and following the standard delay-spread models.

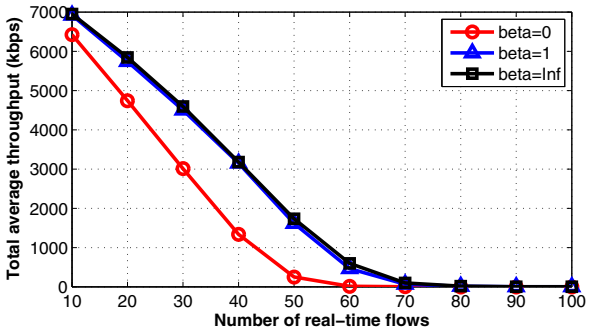
In each simulation trial, we assume that flows have been admitted by the system and the simulation runs for 20000ms. The real-time voice traffic is assumed to be VoIP that periodically generate packets of fixed size, but the inter arrival time has a distribution because each packet can undergo different network situation while passing through the networks from the source to destination. The VoIP traffic is based on G.711 codec standard and generates each VoIP packet every 20 ms, with 160-byte data, 12-byte RTP header, 8-byte UDP header and 20-byte IP header, which results in 200 bytes per MAC layer packet, 80 Kbps data rate. Real-time video streaming traffic has more bursty nature because packet size can be different according to the codec rate such as MPEG-FGS. We assume that the non real-time traffic has no explicit QoS requirements, which means that the traffic is served in a best effort manner.

After we discuss the characteristics of the *beta deadline parameter* in terms of packet loss ratio, sum throughput of the non real-time traffic and response to traffic bursts, we will present traffic class prioritization performance of the proposed scheduling framework.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

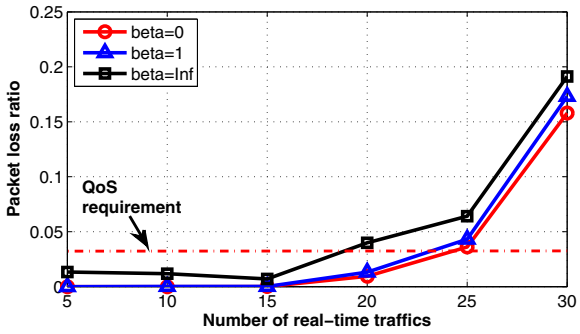


(a) Packet loss ratio vs. number of RT flows

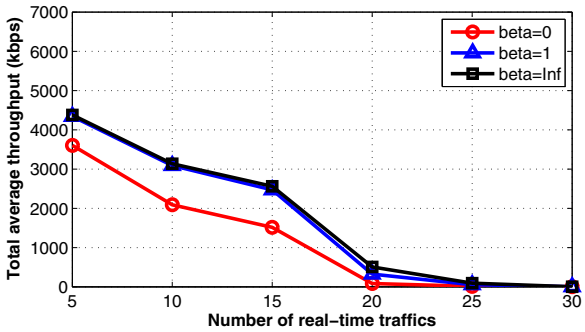


(b) Total average throughput of NRT flows vs. number of RT flows

Fig. 3. Beta deadline parameter characteristics of VoIP traffic



(a) Packet loss ratio vs. number of RT flows



(b) Total average throughput of NRT flows vs. Number of RT flows

Fig. 4. Beta deadline parameter characteristics of MPEG traffic

TABLE II
SUMMARY OF CHARACTERISTICS OF VOICE TRAFFIC

Parameter	Value
Packet inter-arrival time	Exp. dist. with mean=20ms
Packet size	200bytes
Packet loss ratio	$<$ 1%
Maximum latency	$<$ 20ms

TABLE III
SUMMARY OF CHARACTERISTICS OF VIDEO TRAFFIC

Parameters	Value
Frame inter-arrival time	Exp. dist. with mean=40ms
Number of packets in a frame	8
Packet inter-arrival time in a frame	2ms
Packet size	Exp. dist. with mean=128bytes
Packet loss ratio	$<$ 3%
Maximum latency	$<$ 150ms

B. Packet Loss Ratio & Throughput Characteristics

To show *beta deadline parameter* characteristics in terms of packet loss ratio and throughput we run the simulation for the different real-time traffic types, VoIP and MPEG. We assume that packet loss occurs only when the packet is not scheduled within the maximum latency time. The distribution of the real-time traffic and the non real-time traffic is 50:50 meaning that there exists the same number of best effort flows as the voice or video flows. We can identify that beta value 0 has the strongest QoS provisioning characteristic because more than 50 real-time voice flows are guaranteed their QoS requirements comparing that only about 30 real-time voice flows can be possible with the infinite beta value in Fig. 3(a). However, if we focus on the achieved total average throughput of non real-time flows in Fig. 3(b), infinite beta value shows largest total average throughput while beta value 0 reveals the poorest performance. And the results when the beta value is 1, average deadline guarantee scheduling, present the happy medium in terms of throughput of the non real-time traffic and packet loss ratio of the real-time traffic. We can also confirm the trend is similar even the traffic is more burst, such as video streaming through the results in Figs. 4(a), 4(b).

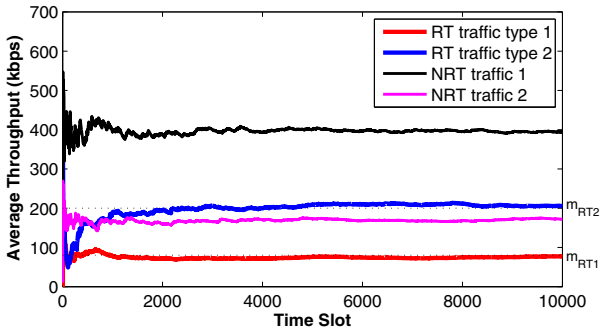
C. Traffic Class Prioritization Performance

The traffic class prioritization performance of the proposed scheduling algorithm is shown in Fig. 5, from which we can see that real-time flows are exactly guaranteed their minimum required data rates regardless of the traffic load. However, the non real-time flows can only attain different shares from the leftover capacity according to their channel quality. In Fig .5, NRT traffic 1 and 2 have the same best effort characteristic but are located differently from the scheduling entity such as base station. Light and heavy traffic load environment is summarized in TABLE IV.

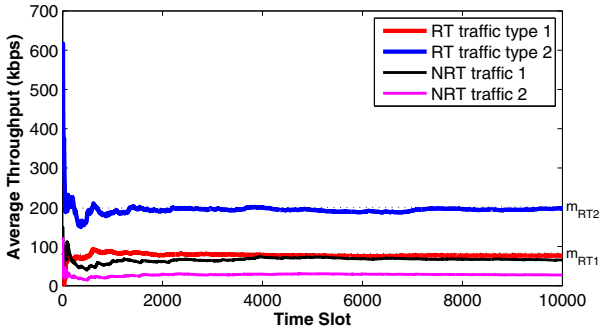
D. Burst Traffic Response Characteristics

When the internet or core network is congested temporarily, base station may get increased traffic load compared to the normal traffic load temporally. We model this burst traffic scenario by changing the traffic load of the real-time voice

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.



(a) Light traffic load



(b) Heavy traffic load

Fig. 5. Traffic class prioritization performance

TABLE IV
SUMMARY OF LIGHT AND HEAVY TRAFFIC LOAD ENVIRONMENT

Traffic load	RT traffic type 1	RT traffic type 2	NRT traffic
Light	10 voice flows	5 video flows	10 BE flows
Heavy	20 voice flows	20 video flows	20 BE flows

traffic. During the 2000 time slot and 3000 time slot, offered traffic rate increases up to 150% of the average traffic rate. During the 7000 time slot and 8000 time slot, offered traffic rate increases to 300% of the average traffic rate. As we expect, beta value 0 has the strongest performance with respect to the packet loss ratio in Fig. 6 due to its prompt adaptation of the traffic load fluctuation.

V. CONCLUSION

Both QoS guarantee and efficient system resource utilization are fundamental research goals in resource-limited wireless networks. The proposed scheduling algorithm satisfies the QoS requirements of the real-time traffic and maximizes the utility of the non real-time traffic while utilizing the system resources efficiently. The simple and well organized *beta deadline parameter* can balances between the QoS provisioning and the opportunity of diversity gain, resulting the effective embodiment of the operating strategy of wireless networks. The characteristics of beta deadline parameter and the performance of proposed scheduling framework is validated through the extensive simulation results. The proposed QoS scheduling framework on the basis of beta deadline parameter

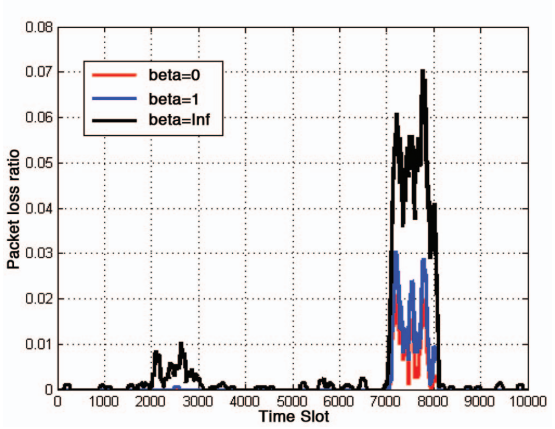


Fig. 6. Burst traffic response

can maximize synergy effects when incorporated with the other non real-time traffic scheduling frameworks.

REFERENCES

[1] Z. Jiang, Y. Ge, and Y. Li, "Max-utility wireless resource management for best-effort traffic," *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 100–111, Jan. 2005.

[2] S. Shakkottai and A. L. Stolyar, "A study of scheduling algorithms for a mixture of real and non-real time data in hdr," Bell Laboratories, Lucent Technologies, Oct. 2000.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[4] A. Jalali, R. Padovani, R. Pankaj, Q. Inc, and C. San Diego, "Data throughput of CDMA-HDR a high efficiency-high data ratepersonal communication wireless system," in *Proc. IEEE VTC*, vol. 3, May 2000, pp. 1854–1858.

[5] A. Khattab and K. Elsayed, "Opportunistic Scheduling of Delay Sensitive Traffic in OFDMA-Based Wireless," in *Proc. IEEE WoWMoM*, June 2006, pp. 279–288.

[6] H. Lee, C. Kim, and S. Chong, "Scheduling and Source Control with Average Queue-Length Control in Cellular Networks," in *Proc. IEEE ICC*, June 2007, pp. 109–114.

[7] G. Li and H. Liu, "Downlink radio resource allocation for multi-cell OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3451–3459, Dec. 2006.

[8] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM system," *J. Select. Areas Commun., IEEE*, vol. 21, no. 2, pp. 171–178, Feb. 2003.

[9] V. Huang and W. Zhuang, "QoS-Oriented Packet Scheduling for Wireless Multimedia CDMA Communications," *IEEE Trans. Mobile Comput.*, pp. 73–85, Jan. 2004.

[10] R. Chipalkatti, J. Jurose, and D. Towsley, "Scheduling policies for real-time and non-real-time traffic in astatistical multiplexer," in *Proc. IEEE INFOCOM*, Apr. 1989, pp. 774–783.

[11] R. Yang, C. Yuan, and K. Yang, "Cross Layer Resource Allocation of Delay Sensitive Service in OFDMA Wireless Systems," in *Proc. IEEE ICCSC*, May 2008, pp. 862–866.

[12] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," *IEEE/ACM Trans. Networking*, vol. 4, no. 3, pp. 375–385, Jun. 1996.

[13] A. Frank, "On Kuhn's Hungarian Method - A tribute from Hungary," *Naval Research Logistics*, vol. 52, no. 1, pp. 2–5, Dec. 2005.

[14] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Networking*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[15] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 76–83, Oct. 2002.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.