EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Adaptive QoS provision for IEEE 802.16e BWA networks based on cross-layer design

Hongtao Zhang[1*], Xiaoxiang Wang[1], ZB Qin[1], GS Kuo[2] and Thomas Michael Bohnert[3]

Abstract

This article proposes an integrated framework for adaptive QoS provision in IEEE 802.16e broadband wireless access networks based on cross-layer design. On one hand, an efficient admission control (AC) algorithm is proposed along with a semi-reservation scheme to guarantee the connection-level QoS. First, to guarantee the service continuity for handoff connections and resource efficiency, our semi-reservation scheme considers both users' handoff probability and average resource consumption together, which effectively avoids resource over-reservation and insufficient reservation. For AC, a new/handoff connection is accepted only when the target cell has enough resource to afford both instantaneous and average resource consumption to meet the average source rate request. On the other hand, a joint resource allocation and packet scheduling scheme is designed to provide packet-level QoS guarantee in term of "*QoS rate*", which can ensure fairness for the services with identical priority level in case of bandwidth shortage. Particularly, an enhanced bandwidth request scheme is designed to reduce unnecessary BR delay and redundant signaling overhead caused by the existing one in IEEE 802.16e, which further improves the packet-level QoS performance and resource efficiency for uplink transmission. Simulation results show that the proposed approach not only balances the tradeoff among connection blocking rate, connection dropping rate, and connection failure rate, but also achieves low mean packet dropping rate (PDR), small deviation of PDR, and low QoS outage rate. Moreover, high resource efficiency is ensured.

Keywords: IEEE 802.16e, QoS model, cross-layer design, adaptive modulation and coding, admission control, resource reservation, bandwidth allocation, scheduling, bandwidth request

1. Introduction

With explosive growth in the data service of Internet and multimedia applications, high-speed and high-quality wireless access is required for providing QoS guarantee for heterogeneous services in future mobile communication systems. As a promising solution for last-mile broadband wireless access (BWA) in metropolitan area, IEEE 802.16d/e [1,2] adopted adaptive modulation and coding (AMC) to maximize the system capacity under the bit error rate (BER) constraint over the error-prone wireless channel [3]. Meanwhile, in the MAC layer, both connection-level and packet-level QoS requirements of heterogeneous services need to be well guaranteed regardless of the channel conditions, and fairness is another important issue to avoid the services with bad channel conditions or low priorities experiencing bandwidth starvation. Particularly, to the uplink transmission in IEEE 802.16e, the fixed/mobile subscriber station (SS) needs to send a bandwidth request (BR) message to base station (BS) for its uplink connection first before data transmission, which introduces additional access delay and signaling overhead for uplink transmission. These characteristics pose great challenge to balance the tradeoff between QoS provision and spectrum efficiency for uplink transmission.

Concerning the service connectivity of the network, the connection-level QoS requirements were achieved through admission control (AC) and resource reservation (RR) [4], whose performance can be evaluated by following metrics: handoff connection dropping rate (CDR), new connection blocking rate (CBR), ongoing connection failure rate (CFR). There are many tradeoffs among these metrics for designing AC and RR schemes. For AC, too stringent restrictions for accepting new/

* Correspondence: htzhang@bupt.edu.cn
[1]Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
Full list of author information is available at the end of the article

handoff connections will reduce the CFR at the cost of deteriorating CBR, CDR, and resource utilization. Even though looser restrictions indicate lower CBR and CDR, too much accepted services may cause system overload, and CFR will greatly increase when the channel condition becomes seriously deteriorated. Since blocking a new connection is more acceptable than dropping a handoff connection from the user viewpoint, performing RR for handoff connections can effectively reduce the CDR. However, over-reservation will deteriorate the CBR and the resource utilization while insufficient reservation cannot achieve prospective CDR target. Therefore, a good AC and RR scheme should well balance these tradeoffs to guarantee the system stability. As for the AC schemes proposed for IEEE 802.16 BWA networks, the authors of [5,6] did not consider the handoff situation, which is a crucial characteristic of IEEE 802.16e. The authors of [7-10] took the handoff requirements in to account regardless of the channel condition. For the general AC schemes proposed in [4,11-18], the time-variant channel conditions were not considered either. In [19], the authors modified the handoff-prioritized AC scheme considering AMC over the unreliable wireless channel, but few QoS-adaptive characteristics were discussed.

The packet-level QoS provision determined the quality of end users experience for multimedia applications [4]. The performance of packet-level QoS provision is evaluated through the metrics including delay, delay jitter, BER, packet loss rate, etc., which is mainly determined by the bandwidth allocation (BA) and scheduling algorithm. In literature, the maximum channel to interference ratio (max C/I) algorithm in [20] was throughput-oriented without QoS consideration, while strict priority queue [21] was QoS-oriented regardless of channel conditions. To better exploit asynchronous variations of channel quality, the authors of [22] gave higher priority to the real-time packets only after their waiting period exceeds the emergency threshold. However, it does not fit well with the bust nature of heterogeneous traffics. Because when large real-time traffics enter the emergency status simultaneously with the bad channel conditions, packet dropping rate (PDR) tends to increase rapidly. Hou et al. [23] introduced the delay constraint into the proportional fairness formulation for QoS provision, but delay is not a proper metric to provide QoS satisfaction and service differentiation for non-real-time traffics. As a variation of modified largest weighted delay first (M-LWDF) [24], the algorithm in [25] considered the channel quality, QoS satisfaction, and service priority for BA. However, the QoS coefficients of various services are not analytically determined. Particularly, the packet-level QoS provision for uplink transmission is also influenced by BR mechanism. Unicast and multicast/broadcast pollings are the primary ways to request bandwidth, while piggyback is an optional way which will not be discussed. The problem of unicast polling is that it introduces constant delay for the delay-sensitive real-time connections. Multicast/broadcast polling provides a contention way to request bandwidth, which causes too much signaling overhead and BR delay for non-real-time services. Lee and Cho [26] reduced the BR delay and signaling overhead for VoIP connections, which did not consider other types of real-time traffics such as MPEG-based multimedia streaming. As for multicast/broadcast polling, the collision probability is a function of the number of BR messages and the contention period size. Oh and Kim [27] and Yan and Kuo [28] proposed two different models to find out the optimal contention period size. The performance of random access for BR was analyzed in [29-31]. Oh and Kim [32] optimized the collision resolution algorithm for BR. However, they cannot eliminate the collisions caused by multicast/broadcast polling because of its contention-based access characteristic. However, the BR delay and the signaling overhead can be further reduced.

Motivated by these observations, we propose an integrated framework for adaptive QoS provision over IEEE 802.16e BWA networks based on cross-layer design, which is considered to be an efficient way to achieve efficient QoS guarantee and network resource management for wireless network [33,34]. Our major contributions are

a) Before accepting a new/handoff connection, the proposed AC algorithm considers whether there is enough bandwidth available to afford its average resource consumption and instantaneous resource consumption for QoS provision through cross-layer design method, which effectively avoids the system overload. So, the proposed AC scheme joint considers the types of service flows (SFs) QoS and MCS, thus embodies the idea of cross-layer design.

b) Our semi-reservation scheme considers both users' handoff probability and average resource consumption together to perform RR, which effectively avoids resource over-reservation and insufficient reservation and ensures well the continuity of handoff connections as well as promises high spectrum efficiency.

c) A joint resource allocation and packet scheduling scheme is designed to guarantee the packet-level QoS in term of "*QoS rate*", thus effectively avoids large real-time data being blocked in deteriorated channel condition. Particularly, when there is not enough bandwidth available to guarantee all "*QoS rate*" constraints, fairness is provided for the services with identical priority level. "*QoS rate*" service model adopts cross-layer design method, since it considers both the bandwidth

requirements in the MAC layer and the channel conditions in the physical layer.

d) An enhanced BR scheme is designed to reduce the unnecessary BR delay and the redundant signaling overhead caused by the existing one in IEEE 802.16e, which further improves the packet-level QoS performance and resource efficiency for uplink transmission.

e) Performing adaptive QoS management to increase or decrease the average source rate based on load status and channel conditions, which enables more users to enter the network, as well as maintains the network stability and high spectral efficiency.

The rest of this article is organized as follows. Section 2 introduces the system model. Section 3 presents the proposed framework for adaptive QoS provision in detail. Section 4 evaluates the system performance through mathematical analysis. Section 5 analyzes the simulation results. Finally, conclusions are made.

2. System model

2.1 QoS-adaptive service model

The MAC layer of IEEE 802.16e is connection-oriented, and a flexible QoS provision framework is designed. Each connection is associated with a unique SF characterizing by a set of QoS parameters such as delay/delay jitter, packet loss rate, minimum reserved rate, maximum sustained rate, etc., and a connection can be created, changed, and deleted through dynamic service addition, dynamic service change, and dynamic service deletion handshake transactions, respectively. Five types of SFs are defined in IEEE 802.16e for QoS differentiation: Unsolicited grant service (UGS), real-time polling service (rtPS), extended rtPS (ErtPS), non-real-time polling service (nrtPS), and best effort (BE). Their priorities from highest to lowest are: UGS, rtPS/ErtPS, nrtPS, and BE. Table 1 lists the characteristics of all SFs.

Considering the influences, i.e., user quantity, channel status (physical layer), service distribution, various QoS restrictions (QoS parameters in application layer), and resource allocation algorithm, that play on the system throughput, a reasonable cross-layer-based mathematical model (*QoS-Adaptive Service Model*) is proposed first to characterize the average system capacity and instantaneous capacity, which is the basis for RR and AC.

Let $C_{m,x,y}$ denote the yth connection belonging to the SF x in subscribe station (SS) m. For UGS, rtPS/ErtPS, nrtPS, and BE, the value of x equals 1, 2, 3, and 4, respectively. In this article, the traffic sources are considered to be rate adaptive, because different coding schemes are provided for multimedia services in application layer. We set $G_{m,x,y}$ service grades for the connection $C_{m,x,y}$. Let $R^{\min}_{m,x,y}$ and $R^{\max}_{m,x,y}$ be the minimum rate and the maximum rate of the connection $C_{m,x,y}$, respectively. For a connection at service grade g, its average required rate for QoS provision can be defined as

$$R^{\text{avg}}_{m,x,y,g} = R^{\min}_{m,x,y} + \frac{(R^{\max}_{m,x,y} - R^{\min}_{m,x,y})(g-1)}{G_{m,x,y} - 1} \quad 1 \le g \le G_{m,x,y} \quad (1)$$

It is obvious that the smaller g indicates lower average required rate for QoS provision, and vice versa.

For the connection $C_{m,x,y}$, $D_{m,x,y}$, $W_{m,x,y}$, $\psi_{m,x,y}$, and $\omega_{m,x,y}$, respectively, denote the tolerable delay, the waiting period of its packets before being transmitted, the packet error rate (PER) during transmission and the tolerable packet loss rate. A packet may be dropped when transmission error happens or its waiting period exceeds the tolerable delay. Thus, Equation 2 must be met to avoid the ongoing connection failure.

$$\Pr\{W_{m,x,y} > D_{m,x,y}\} + \psi_{m,x,y} \le \omega_{m,x,y} \quad (2)$$

In the following section, we will find that the PER can be guaranteed through selecting proper modulation and coding scheme (MCS) based on the SINR knowledge. Thus, the resource allocation and scheduling algorithm should guarantee the maximum delay for a given outage probability. Particularly, reducing BR delay for uplink connections can help for reducing the PDR caused by delay variation. However, because of the burst nature of heterogeneous traffics, $R^{\text{avg}}_{m,x,y,g}$ cannot accurately reflect the instantaneous rate requirements to provide QoS guarantee for the connection $C_{m,x,y}$. Accordingly, based on cross-layer method, the term "*QoS rate*" is defined in Equation 3 for packet-level QoS provision (upper-layer),

Table 1 SF characteristics

SF	Traffic type	QoS constraint
UGS	Constant bit rate (CBR-based) services (e.g., the leased line E1/T1, VoIP without compression)	Stringent requirements on data rate, delay/delay jitter and packet loss rate
rtPS	Real-time variable bit rate (VBR-based) services (e.g., mpeg-based video conference and multimedia streaming)	Strict delay and packet loss rate requirements
ErtPS	Real-time VBR-based services (e.g., VoIP without compression)	
nrtPS	Non-real-time VBR-based services (e.g., FTP)	Minimum reserved rate and stringent packet loss rate requirements
BE	BE services (e.g., HTTP, E-mail)	Packet loss rate should be maintained

which considers both delay constraint and the minimum/maximum rate constraints (data link layer) together.

$$R^{q}_{m,x,y} = \text{Min}\{\text{Min}\{R_{m,x,y}, R^{\max}_{m,x,y}\}, \quad \text{Max}\{R^{\text{em}}_{m,x,y}, R^{\min}_{m,x,y}\}\} \quad (3)$$

where $R^{q}_{m,x,y}$, $R_{m,x,y}$, and $R^{\text{em}}_{m,x,y}$ are the "*QoS rate*", the required rate to transmit the buffer data, and the rate to transmit the emergency data for the connection $C_{m,x,y}$, respectively. The emergency data are the data whose waiting periods exceed the tunable delay threshold $\xi_{m,x,y}$ ($0 < \xi_{m,x,y} < D_{m,x,y}$). Since both deteriorated channel condition and increased source rate may cause higher $\Pr\{W_{m,x,y} > D_{m,x,y}\}$, smaller $\xi_{m,x,y}$ should be considered, and vice versa. And the "*non-QoS rate*" of the connection $C_{m,x,y}$ can be defined as $R^{\text{nq}}_{m,x,y} = R_{m,x,y} - R^{q}_{m,x,y}$. Accordingly, we have $R^{\text{em}}_{m,x,y} = 0$ for the delay insensitive nrtPS/BE connections, $R^{q}_{m,x,y} = 0$ for the BE connections without minimum rate requirement, $R^{q}_{m,x,y} = R^{\max}_{m,x,y} = R^{\min}_{m,x,y}$ and $R^{\text{nq}}_{m,x,y} = 0$ for UGS connections with fixed rate requirements.

2.2 Link adaptation model

This article considers the PHY layer of IEEE 802.16e BWA networks combining WirelessMAN-OFDM with AMC together for optimizing the system performance over the error-prone wireless channel. As a TDMA-based PHY technology, each frame of WirelessMAN-OFDM contains many transmission bursts from/to different SSs. The data rate and coding overhead for each burst are different, because different MCSs are chosen for the SSs for adapting to various detected signal-to-noise ratios (SNRs), and to meet the target BER accordingly. Since M-QAM modulation provides high spectrum efficiency while convolutional codes (CC) with bit interleaved coded modulation have strong forward error protection capability, they are chosen to form MCS compositions. The entire SINR range is divided into $K + 1$ non-overlapping consecutive partitions by the SINR boundary Γ_k ($1 \le k \le K$), and $\Gamma_1 < \Gamma_2 < ... < \Gamma_K = \infty$. If the SINR is in the range of $(\Gamma_k, \Gamma_{k+1}]$, MCS k is adopted. Particularly, because of unacceptable transmission error, no data are transmitted if the SINR is less than Γ_1. The MCS employed in this article is listed in Table 2. If SS m adopts MCS k, its average PER can be deduced as

$$\psi^{k}_{m} = \sum_{l=\eta_k}^{L} \binom{L}{l} (\varepsilon_m)^{l} (1-\varepsilon_m)^{L-l} \quad (4)$$

where L is the average packet length, η_k is the number of error bits can be corrected by MCS k, and ε_m is the BER constraint of SS m.

Adopting MCS k, the data rate from MAC layer viewpoint can be calculated as

$$MR_k = B\Omega_k \lfloor PR_k CR_k / \Omega_k \rfloor \quad (5)$$

where B, PR_k, Ω_k, and CR_k, respectively, denote the channel bandwidth, PHY transmission rate, the modulation level, and the CC code rate when MCS k is adopted. It is noted that the modulation levels of QPSK, QAM16, and QAM64 are 2, 4 and 6, respectively.

To analyze the system capacity over the time-variant wireless channel, we assume that both the path loss and shadowing are compensated by dynamically adjusting the transmission power. Thus, only the small-scale fading need to be considered. For SS m, the probability density function of its SNR γ_m under the Rayleigh fading environment is

$$\Pr(\gamma_m) = \frac{1}{\bar{\gamma}_m} \exp\left(-\frac{\gamma_m}{\bar{\gamma}_m}\right) \quad (6)$$

where $\bar{\gamma}_m$ is the average SNR of SS m. Accordingly, the probability of an SS adopting MCS k for transmission can be deduced as

$$\mathbb{P}_m(k) = \int_{\Gamma_k}^{\Gamma_{k+1}} \Pr(\gamma) d\gamma = \exp\left(-\frac{\Gamma_k}{\bar{\gamma}_m}\right) - \exp\left(-\frac{\Gamma_{k+1}}{\bar{\gamma}_m}\right) \quad (7)$$

The average resource consumption (transmission time) for transmitting one bit can be deduced as

$$\mathbb{C}^{avg}_{m} = \sum_{k=1}^{K} \mathbb{P}_m(k) / MR_k \quad (8)$$

Based on the QoS-adaptive characteristics in the MAC layer and the average resource consumption (transmission time) per bit in the PHY layer (Equation 8), we proceed to investigate cross-layer design for bandwidth resource management in the following section.

3. Cross-layer design for QoS-adaptive resource management

In the point-to-multipoint (PMP) mode of IEEE 802.16e BWA networks, BS is designed as a coordinator to perform QoS-adaptive resource management for its subordinate fixed/mobile SSs. The proposed adaptive QoS provision framework and the interaction between BS and SS are shown in Figure 1. At the connection level, the admission controller in BS restricts the number of new/handoff connections entering the target cell to avoid system overload, which ensures low CFR of ongoing connection. In addition, the RR executes the semi-reservation algorithm, which not only guarantees the service continuity for handoff connections, but also achieves high resource efficiency, that is because it

Table 2 Modulation and coding schemes

K	Modulation	CC code rate CR_k	PHY transmission rate PR_k(bits/s/Hz)	SINR (dB) for BER ≤ 10^{-6}
1	QPSK	1/2	1.00	4.65
2	QPSK	2/3	1.33	6.49
3	QPSK	3/4	1.50	7.45
4	QAM16	1/2	2.00	10.93
5	QAM16	2/3	2.66	12.71
6	QAM16	3/4	3.00	14.02
7	QAM64	2/3	4.00	18.50
8	QAM64	3/4	4.5	19.88
9	QAM64	7/8	5.25	21.94

effectively avoids resource over-reservation and insufficient reservation. Particularly, the SF managers in SS and BS communicate with each other to maintain the connections' survival as well as perform adaptive QoS adjustment. At the packet level, the resource dispenser in BS takes SS as the basic unit to perform resource allocation through cross-layer design idea, which considers both the "*QoS rate*" constraints in the MAC layer and the channel conditions in the PHY layer. The resource allocation result for downlink transmission is reflected in DL-MAP message, while the one for uplink transmission is figured out in UL-MAP message. Using the granted bandwidth for each SS, schedulers in BS and SS schedule the downlink and uplink data for transmission, respectively. Specifically, when an SS has data to send in the uplink, it needs to send a BR message to BS first. A BR generator is designed to execute the proposed BR scheme, which can help to reduce the BR delay and signaling overhead. Since the difference between uplink and downlink transmission mainly lies in whether a connection needs to request bandwidth before data transmission, for simplicity, we only discuss the uplink case for QoS provision in this article. Following sections will describe our proposed approach (PA) in detail.

3.1 The estimation of RR

We extend the probabilistic resource estimation and semi-reservation scheme [35] for reasonable RR considering the time variant channel conditions. For a mobile

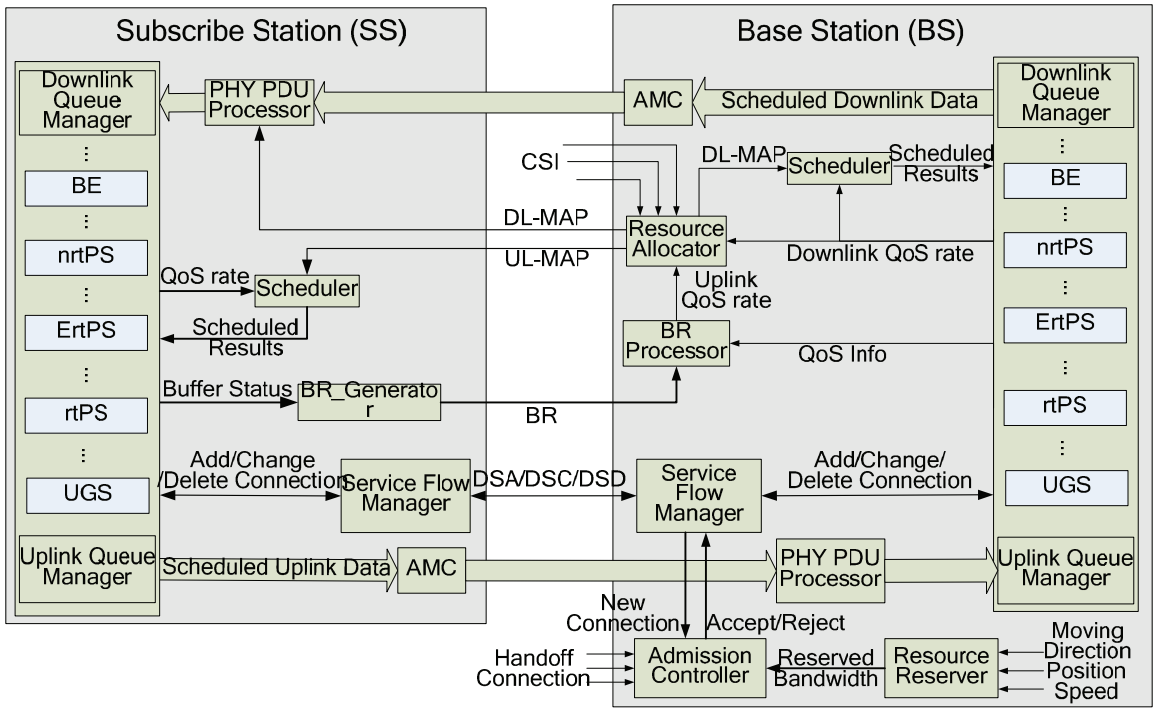

Figure 1 Proposed adaptive QoS provision framework.

SS m managed by cell u, $H_{m,u,v}$ denotes the handoff probability from cell u to cell v, which can be calculated based on the current position, as well as the predicted moving speed and direction of mobile SS m. Let NC_u denote the collection of the neighboring cells of cell u. We have

$$\sum_{v \in NC_u} H_{m,u,v} + H_{m,u,u} = 1 \tag{9}$$

To reduce unnecessary RR, *reservation threshold* Δ is defined. Reservations are made only for the mobile SSs with handoff probabilities larger than Δ. Let $Y_{m,x}$ be the number of the connections belonging to SF x in mobile SS m. Set $z_{m,u} = 1$, if SS m is in cell u. Otherwise, $z_{m,u} = 0$. Meanwhile, set $\Re_{m,x,y} = R^{\text{avg}}_{m,x,y,g}$ for the connection $C_{m,x,y}$ at service grade g. Suppose there are M SSs distributed in the whole network. In cell v, if $H_{m,u,v} > \Delta$, the average reserved bandwidth for the connections belonging to SF x can be deduced as

$$RS_{v,x} = \sum_{u \in NC_v} \sum_{m}^{M} \sum_{y=1}^{Y_{m,x}} z_{m,u} H_{m,u,v} \Re_{m,x,y} \mathbb{C}^{\text{avg}}_{m} \tag{10}$$

It is noted that in the above equation, $RS_{v,x}$ is the bandwidth co-reserved for the connections belonging to SF x other than for a specific connection or mobile SS. Accordingly, the total reserved bandwidth in cell v can be deduced as $RS_v = \sum_{x=1}^{4} RS_{v,x}$.

3.2 Admission control

In this section, we discuss AC considering both instantaneous resource consumption and average resource consumption. Let $\text{AS}_{v,x}$ and $\text{PS}_{v,x}$ denote the instantaneous and average resource consumption of the connections belonging to SF x in cell v, respectively. k_m is serial number of the selected MCS based on the instantaneous SNR of SS m. We have

$$\begin{cases} \text{AS}_{v,x} = \sum_{m=1}^{M} \sum_{y=1}^{Y_{m,x}} z_{m,v} \Re_{m,x,y} \mathbb{C}^{\text{avg}}_{m} \\ \text{PS}_{v,x} = \sum_{m=1}^{M} \sum_{y=1}^{Y_{m,x}} z_{m,v} \Re_{m,x,y} / MR_{k_m} \end{cases} \tag{11}$$

Thus, the total average and instantaneous resource consumption in cell v can be calculated as $\text{AS}_v = \sum_{x=1}^{4} \text{AS}_{v,x}$ and $\text{PS}_v = \sum_{x=1}^{4} \text{PS}_{v,x}$, respectively.

In case of bandwidth shortage, more new/handoff real-time connections can be accepted by decreasing the source rate of the ongoing connections which are not prioritize over them. Accordingly, the average resource $\overline{\text{PS}_{v,x}}$ and instantaneous resource $\overline{\text{PS}_{v,x}}$ must be reserved for ongoing connections after decreasing source rate for new/handoff connections belonging to SF x. Actually, since we satisfy bandwidth requirements (QoS) in upper-layer through source rate compression, i.e., decreasing transmission rate in data link layer via MCS, this proposed scheme embodies the idea of cross-layer design. In order to guarantee the minimum QoS requirements of ongoing connections, the average resource $\overline{\text{AS}_{v,x}}$ and instantaneous resource $\overline{\text{PS}_{v,x}}$ can be deduced as

$$\begin{cases} \overline{\text{AS}_{v,x}} = \text{AS}_v - \sum_{s=x}^{4} \text{AS}_{v,s} + \sum_{s=x}^{4} \sum_{m=1}^{M} \sum_{y=1}^{Y_{m,x}} z_{m,v} R^{\min}_{m,s,y} \mathbb{C}^{\text{avg}}_{m} \\ \overline{\text{PS}_{v,x}} = \text{PS}_v - \sum_{s=x}^{4} \text{PS}_{v,s} + \sum_{s=x}^{4} \sum_{m=1}^{M} \sum_{y=1}^{Y_{m,x}} z_{m,v} R^{\min}_{m,s,y} / MR_{k_m} \end{cases} \tag{12}$$

Since blocking a new connection is more acceptable than dropping an ongoing connection from the user viewpoint, the bandwidth reserved for handoff connection cannot be used for accepting new connection. Let TS_v be the total available bandwidth in cell v. For a new connection meeting both inequalities in Equation 13, it will be accepted at its desired average source rate without source rate compression for other connections. In case of bandwidth shortage, a new connection is accepted at its minimum rate if the constraints in Equation 14 are met, which may causes the source rates of other connections being decreased. If neither Equations 13 nor 14 is met, the new connection will be rejected.

$$\begin{cases} \Re_{m,x,\text{new}} \mathbb{C}^{\text{avg}}_{m} \le \text{TS}_v - \text{AS}_v - \text{RS}_v \\ \Re_{m,x,\text{new}} / \text{MR}_{k_m} \le \text{TS}_v - \text{PS}_v - \text{RS}_v \end{cases} \tag{13}$$

$$\begin{cases} R^{\min}_{m,x,\text{new}} \mathbb{C}^{\text{avg}}_{m} \le \text{TS}_v - \overline{\text{AS}_{v,x}} - \text{RS}_v \\ R^{\min}_{m,x,\text{new}} / \text{MR}_{k_m} \le \text{TS}_v - \overline{\text{PS}_{v,x}} - \text{RS}_v \end{cases} \tag{14}$$

In our scheme, the handoff connection with higher priority may preempt the bandwidth reserved for the lower priority ones. Thus, the reserved bandwidth, which cannot be used by the handoff connections belonging to SF x, is $\overline{\text{RS}_{v,x}} = \text{RS}_v - \sum_{s=x}^{4} \text{RS}_{v,s}$. A handoff connection is accepted at its desired source rate if both inequalities in Equation 15 are met, which neither preempt the reserved bandwidth of the handoff connections belonging to other SFs, nor compress the sources rate of the ongoing connections. When there is not enough resource available, a handoff connection is accepted at its minimum rate. In this case, either reserved bandwidth preemption or the source rate degradation may happen. If neither Equations 15 nor 16 are met, the handoff connection will be dropped.

$$\begin{cases} \Re_{m,x,ho}\mathbb{C}_m^{\text{avg}} \le \text{TS}_v - \text{AS}_v - (\text{RS}_v - \text{RS}_{v,x}) \\ \Re_{m,x,ho}/\text{MR}_{k_m} \le \text{TS}_v - \text{PS}_v - (\text{RS}_v - \text{RS}_{v,x}) \end{cases} \quad (15)$$

$$\begin{cases} R_{m,x,ho}^{\min}\mathbb{C}_m^{\text{avg}} \le \text{TS}_v - \overline{\text{AS}_{v,x}} - \overline{\text{RS}_{v,x}} \\ R_{m,x,ho}^{\min}/\text{MR}_{k_m} \le \text{TS}_v - \overline{\text{PS}_{v,x}} - \overline{\text{RS}_{v,x}} \end{cases} \quad (16)$$

3.3 Adaptive QoS management

Accepting more new/handoff connection in case of bandwidth shortage is not the only reason to perform source rate compression. Since AC can keep $\text{AS}_v \le \text{TS}_v$ for cell v, if the available bandwidth cannot afford all ongoing connections' average source rate and "*QoS rate*" requirements because of the deteriorated channel conditions, source rate compression will be performed to keep the system stable. In this case, either inequality in Equation 17 is met.

$$\begin{cases} \text{PS}_v > \text{TS}_v \\ \sum_{m=1}^{M}\sum_{x=1}^{4}\sum_{\gamma=1}^{Y_{m,x}} z_{m,v}R_{m,x,\gamma}^{q}/\text{MR}_{k_m} > \text{TS}_v \end{cases} \quad (17)$$

For source rate compression, the connections with lower priority are chose first. Among the connections with identical priority level, the connection whose master SS has the worst channel condition is chosen first. The selected connection can adapt to any coding scheme producing lower average source rate, and least number of degraded connections should be selected to reduce the signaling overhead.

If all "*QoS rate*" constraints of ongoing connections are guaranteed and there is still bandwidth left unused exempting the reserved bandwidth, we will increase ongoing connections' average source rate to improve the resource utilization and the service quality. Among the connections whose average source rates have been compressed, the one whose master SS has best channel condition will be chosen first. Then, for other connections, the one with highest priority level among the ones with best channel condition is chosen. The selected connection can adapt to its highest average source rate for reducing the signaling overhead as well as improving the system throughput.

3.4 Enhanced BR scheme

The term "QoS rate" is defined in Equation 3 to reflect the time-variant QoS requirement of the service because of its bursty characteristics. Based on this definition, a joint resource allocation and scheduling algorithm is designed to provide QoS guarantee based on "QoS rate" as well as fairness for the services with identical priority level in case of bandwidth shortage. Specifically, an enhanced BR mechanism is proposed, which reduces the number of bandwidth request messages by aggregating the nrtPS/BE connections in the same SS as one basic BR unit, as well as replaces the reactive unicast polling and multicast/broadcast polling with proactive unicast polling to reduce the BR delay and signaling overhead.

We enhance the BR scheme for IEEE 802.16e BWA networks in the following aspects:

a) SS requests bandwidth only using unicast polling opportunity, which avoids the BR collisions caused by multicast/broadcast polling.

b) Each uplink rtPS/ErtPS connection is taken as an individual BR *unit* (BRU) because of the stringent delay requirement, while all uplink nrtPS or BE connections in the same SS are aggregated as a *BRU* to reduce the signaling overhead for unicast polling.

c) The uplink protocol data units (PDUs) have two statuses: *transmission-preparing* (*tp*) and *transmission-ready* (*tr*). The incoming uplink data are packed into the PDUs in *tp* status first. Once SS requests bandwidth for a *BRU*, the BR message takes the aggregated bandwidth requirement for all its PDUs to BS, and the PDUs of the *BRU* in *tp* status are transited to *tr* status accordingly.

d) The reserved bit in generic MAC header is defined as *unicast polling index* (UPI). When SS needs to be polled, UPI is set to 1; otherwise, it is set to 0.

Based on the *BRU* definition in (b), in BS, $C_{m,x,\gamma}$ can also be used to denote the corresponding rtPS/ErtPS *BRU*, while $C_{m,x,-1}$ is used to represent the nrtPS/BE *BRU* in SS m. For an uplink *BRU* $C_{m,x,\gamma}$, $R_{m,x,\gamma}^{tp}$ and $R_{m,x,\gamma}^{tr}$ represent the bandwidth requirement of its PDUs in *tp* and *tr* statuses, respectively. It is noted that only the PDUs in *tr* status can be transmitted out when there is uplink bandwidth available. To obtain the "*QoS rate*" of each uplink *BRU* in BS, we have $R_{m,x,-1}^{\min} = \sum_{\gamma=1}^{Y_{m,x}} R_{m,x,\gamma}^{\min}$, $R_{m,x,-1}^{\max} = \sum_{\gamma=1}^{Y_{m,x}} R_{m,x,\gamma}^{\max}$ and $R_{m,x,\gamma} = R_{m,x,\gamma}^{tp}$. Based on the definition in Equation 3, in BS, the uplink "*QoS rate*" of UGS/rtPS/ErtPS in SS m can be defined as $R_{m,x}^{q} = \sum_{\gamma=1}^{Y_{m,x}} R_{m,x,\gamma}^{q}$, while the one for nrtPS/BE is $R_{m,x}^{q} = R_{m,x,-1}^{q}$. The emergency rate of real-time SF x in SS m meets $R_{m,x}^{\text{em}} = \sum_{\gamma=1}^{Y_{m,x}} R_{m,x,\gamma}^{\text{em}}$, and the "*non-QoS rate*" of SS m can be defined as

$$R_m^{\text{nq}} = \sum_{\gamma=1}^{Y_{m,2}} R_{m,2,\gamma}^{\text{nq}} + \sum_{x=3}^{4}\sum_{\gamma=1}^{Y_{m,2}} R_{m,x,\gamma}^{tp} - \sum_{x=3}^{4} R_{m,x,-1}^{q} \quad (18)$$

Let $\eta_{m,x,\gamma}$ be the tunable variable for the *BRU* $C_{m,x,\gamma}$ to set UPI. If SS requests bandwidth for a *BRU* once

new data come in, the bandwidth requirement can be reflected to BS in the shortest time at the cost of highest signaling overhead. We define the following rules to balance the tradeoff between the two issues: (1) when there is uplink bandwidth available, SS first requests bandwidth for the *BRU*s with expired unicast polling timer, then for the *BRU*s which UPIs have been set for; (2) if there are data PDU to be sent out, the SS sets UPI for *BRU* based on Equations 19 and 20 for rtPS/ErtPS and nrtPS/BE, respectively. Figure 2 depicts the operations of the enhanced BR scheme in SS.

$$\max\{R^{tr}_{m,x,\gamma}, \eta_{m,x,\gamma}\} \le R^{tp}_{m,x,\gamma} \tag{19}$$

$$\max\left\{\sum_{\gamma=1}^{Y_{m,x}} R^{tr}_{m,x,\gamma}/Y_{m,x}, \eta_{m,x,-1}\right\} < \sum_{\gamma=1}^{Y_{m,x}} R^{tp}_{m,x,\gamma} \tag{20}$$

Once BS receives an uplink PDU with UPI equaling one, in next frame, it will grant a unicast polling opportunity to the SS which sends the PDU. In addition, the SS whose *BRU* has expired unicast polling timer will also be granted a unicast polling opportunity in next frame. The operations of the enhanced BR scheme in BS are shown in Figure 3.

3.5 Joint BA and scheduling

BS follows strict priority to process the "*QoS rate*" requirements for its subordinated SSs, and the detailed resource allocation algorithm is designed based on Equation 21.

$$X_{\max} = \arg\max_{X} \sum_{x=1}^{X} \sum_{m=1}^{M} z_{m,v} R^{q}_{m,x}/MR_{k_m} \le TS_v \tag{21}$$

The channel condition is seriously deteriorated or the real-time traffic is boosted when $X_{\max} < 2$, which cause the available bandwidth cannot satisfy all "*QoS rate*" requirements of the real-time SF $X_{\max}$. In this case, BS will prior guarantee the emergency rate requirements other than "*QoS rate*" requirements. Even worse, if the available bandwidth cannot afford their emergency rate requirements, packet loss may happen. All SSs should share the packet loss to avoid the SSs with deteriorated channel condition suffering from more serious QoS degradation. So, BS chooses to serve the SS with the lowest satisfaction for emergency rate in recent *S* fames first.

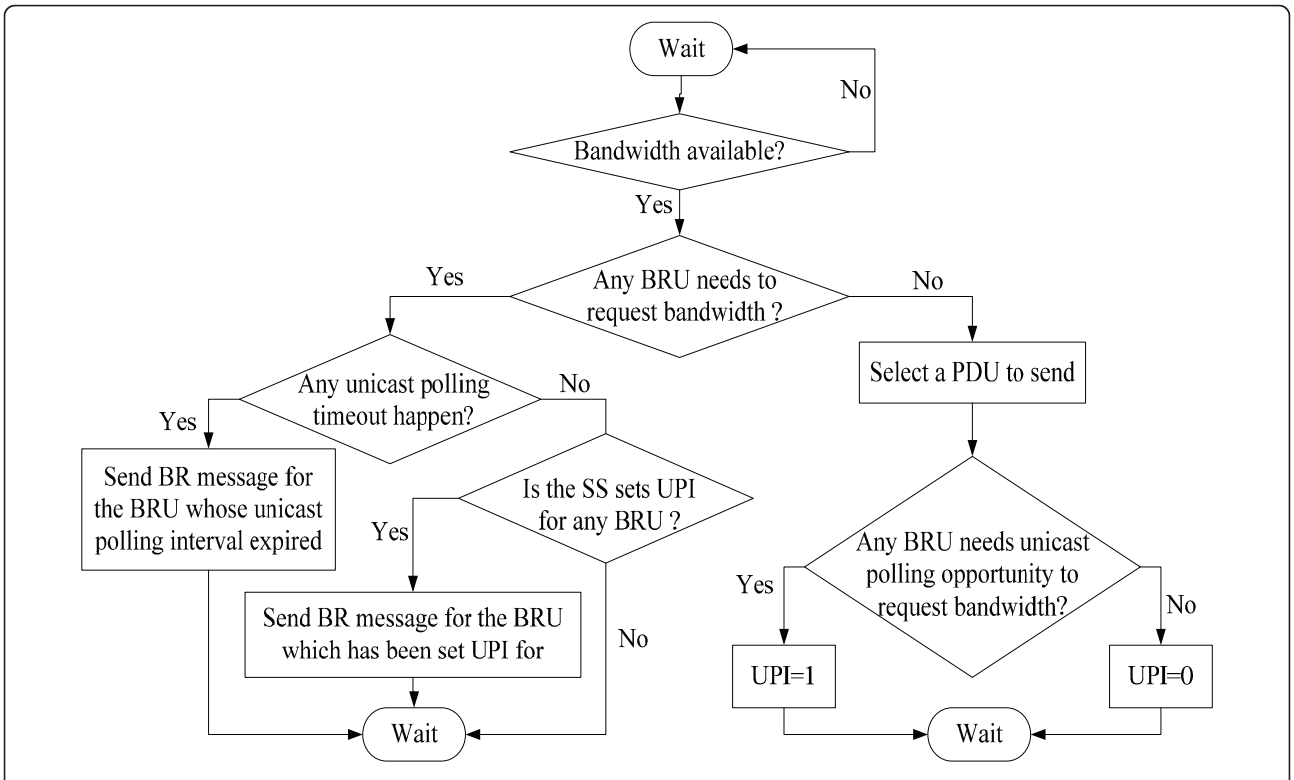

Figure 2 The operations of the enhanced BR scheme in SS.

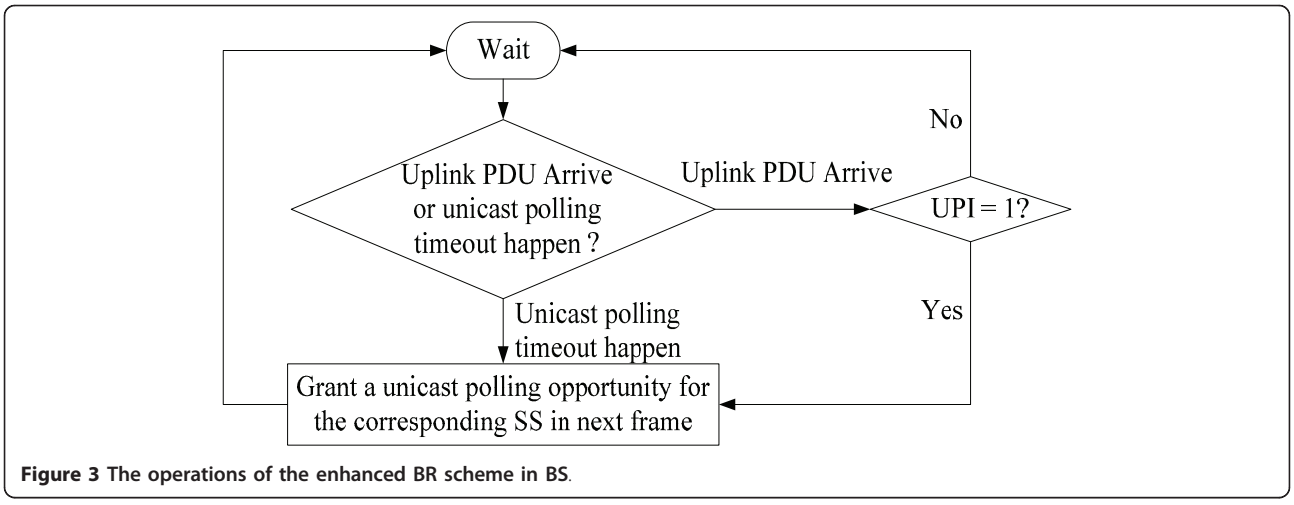

Figure 3 The operations of the enhanced BR scheme in BS.

$X_{\max} = 2$ indicates that the available bandwidth cannot satisfy all "*QoS rate*" requirements of nrtPS. Let $Th_{v,3}$ be the MAC throughput of the available bandwidth $TS_{v,3}$ for nrtPS, thus $G^q_{m,v,3}$ satisfies (22).

$$\sum_{m=1}^{M} G^q_{m,v,3} = TS_{v,3} \tag{22}$$

To provide weighted fairness for the "*QoS rate*" requirements of nrtPS connections from MAC viewpoint, Equation 23 can be used to deduce the bandwidth granted to each SS for nrtPS connections.

$$G^q_{m,v,3} = W^q_{m,v,3} * Th_{v,3}/MR_{k_m} \tag{23}$$

where $W^q_{m,v,3} = z_{m,v}R^q_{m,3}/\sum_{m=1}^{M} z_{m,v}R^q_{m,3}$.

Using the granted bandwidth of each SS, we perform packet scheduling for its connections considering "*QoS rate*" requirements first, and the scheduling rules are defined as: (1) for the connections belonging to different SF types, their packets are scheduled following order of strict priority; (2) for the connections belonging to the same SF type, the connection whose head-of-line packet has the longest waiting time will be served first. After QoS provision, SS applies round robin [36] to schedule packets based on the "*non-QoS rate*" requirements of its connections.

4 Mathematical analysis

In this section, we perform mathematical analysis for the PA from following viewpoints: connection-level QoS performance, queuing performance, and BR efficiency enhancement.

4.1 Connection-level QoS performance analysis

To simplify theoretical analysis for connection-level QoS provision, we assume (1) all connections in cell v belong to SF x, and they have the same minimum rate requirement $R_x^{\min}$; (2) the average SNR $\bar{\gamma}_m$ is identical for all SSs. So, $\mathbb{P}_m(k)$ in Equation 7 and $\mathbb{C}_m^{avg}$ in Equation 8 can be simplified as $\mathbb{P}(k)$ and $\mathbb{C}^{\text{avg}}$, respectively. Set $M_v = \sum_{m=1}^{M} z_{m,v}$. The probability of s connections adopting MCS k for data transmission can be deduced as

$$\tilde{\mathbb{P}}(k, s, M_v) = \binom{M_v}{s} (\mathbb{P}(k))^s (1 - \mathbb{P}(k))^{M_v - s} \tag{24}$$

Accordingly, the characteristic function of the above equation is

$$\varphi(k, s, M_v, z) = \sum_{s=0}^{M_v} \tilde{\mathbb{P}}(k, s, M_v) z^s = \left(1 - \mathbb{P}(k) + z\mathbb{P}(k)\right)^{M_v} \tag{25}$$

The average number of the connections adopting MCS k for data transmission is

$$\Psi(k, M_v) = \sum_{s=0}^{M_v} s * \tilde{\mathbb{P}}(k, s, M_v) \tag{26}$$

Suppose all the connections in cell v are at their lowest average source rate, the average resource consumption in cell v can be calculated as

$$\begin{aligned} \alpha_v &= \sum_{k=1}^{K} \Psi(k, M_v) * R_x^{\min}/MC_k = \sum_{k=1}^{K} \frac{d\varphi(k, s, M_v, z)}{dz}|_{z=1} * R_x^{\min}/MC_k \\ &= \sum_{k=1}^{K} M_v * R_x^{\min}\mathbb{P}_m(k)/MC_k = M_v R_x^{\min}\mathbb{C}^{\text{avg}} \end{aligned} \tag{27}$$

To investigate the tradeoff among CBR, CDR, and CFR under the time-variant channel conditions, we study two extreme cases. One case is that all connections are in the best channel conditions, and we have $PS_v = M_v * R_x^{\min}/MR_K$. If $PS_v = TS_v$, the value of M_v is maximized, which implies that lower CBR and CDR are ensured. However, once the channel condition gets worse, $PS_v > TS_v$ will be met. Since no rate compression can be performed, many connection may fail, which will result in system unstability. The average CFR in this case can be calculated as

$$\mathrm{CFR}_v = 1 - TS_v/(M_v R_x^{\min} \mathbb{C}_m^{\mathrm{avg}}) = 1 - 1/MR_K \mathbb{C}_m^{\mathrm{avg}} \quad (28)$$

The other case is that only lowest transmission rate is available because of the deteriorated channel condition. When all bandwidth resources are used up, we have $TS_v = M_v * R_x^{\min}/MR_1$. In this case, it is obvious that the lowest CFR is available at the cost of highest CDR and CBR, because from average viewpoint, there still be a lot of new/handoff connections can be accepted by the system, the number of which can be deduced as

$$\beta_v = (TS_v - \alpha_v)/R_x^{\min} = M_v(1/MR_1 - \mathbb{C}_m^{\mathrm{avg}}) \quad (29)$$

4.2 Queuing performance analysis

In this section, we analyze the queuing performance for the network under the saturated status, in which all the available bandwidths are used up to guarantee all ongoing connections' average rate requirement for QoS guarantee. Since UGS connections always get fixed bandwidth for data transmission without BR, we only discuss other types of connections here. The uplink data access delay consists of BR delay and scheduling delay. By setting the BR delay equals 0, the analysis result for uplink transmission can be extend to the downlink transmission as well.

For an non-UGS uplink connection $C_{m,x,y}$, suppose its data arrival follows Poisson process with rate $\lambda_{m,x,y}$ packets per second, and the average length of the packet is $L_{m,x,y}$. We have $\lambda_{m,x,-1} = \sum_{y=1}^{Y_{m,x}} \lambda_{m,x,y}$ for nrtPS/BE *BRU*. Due to the effect of Equations 19 and 20, the average rate of the uplink data transmitting from *tr* status to *tp* status is $q_{m,x,y}$ bits per second. Therefore, the uplink transmission process of a BRU can be formulated as a twice queuing problem shown in Figure 4, which can be depicted by the two-dimensional Markov model shown in Figure 4. The steady-state equation in Figure 5 is obtained as Equation 30, in which $\mu_{m,x,y} = q_{m,x,y}/L_{m,x,y}$ and $I_{m,x,y} = \Re_{m,x,y}/L_{m,x,y}$.

$$\begin{cases} \lambda_{m,x,y} p_{0,0} = \iota_{m,x,y} p_{0,1} & \\ (\lambda_{m,x,y} + \mu_{m,x,y}) p_{r,0} = \iota_{m,x,y} p_{r,1} + \lambda_{m,x,y} p_{r-1,0} & r \geq 1 \\ (\lambda_{m,x,y} + \iota_{m,x,y}) p_{0,s} = \iota_{m,x,y} p_{0,s+1} + \mu_{m,x,y} p_{1,s-1} & s \geq 1 \\ (\lambda_{m,x,y} + \mu_{m,x,y} + \iota_{m,x,y}) p_{r,s} = \iota_{m,x,y} p_{r,s+1} + \mu_{m,x,y} p_{r+1,s-1} + \lambda_{m,x,y} p_{r-1,s} & r \geq 1 \& s \geq 1 \\ \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} p_{r,s} = 1 & \end{cases} \quad (30)$$

Using recursive algorithm, the steady-state probability for each state can be obtained as

$$p_{r,s} = (\rho_{m,x,y}^{tp})^r (\rho_{m,x,y}^{tr})^s (1 - \rho_{m,x,y}^{tp})(1 - \rho_{m,x,y}^{tr}) \quad (31)$$

where $\rho_{m,x,y}^{tp} = \lambda_{m,x,y}/\mu_{m,x,y}$ and $\rho_{m,x,y}^{tr} = \lambda_{m,x,y}/\iota_{m,x,y}$. Based on Equation 31, the average queuing length in *tp* buffer and *tr* buffer can be deduced as

$$\begin{cases} \bar{r} = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} r p_{r,s} = \rho_{m,x,y}^{tp}/(1 - \rho_{m,x,y}^{tp}) \\ \bar{s} = \sum_{s=0}^{\infty} \sum_{r=0}^{\infty} s p_{r,s} = \rho_{m,x,y}^{tr}/(1 - \rho_{m,x,y}^{tr}) \end{cases} \quad (32)$$

And the queuing delay in *tp* buffer and *tr* buffer can be deduced as

$$\begin{cases} \overline{D_{m,x,y}^{tp}} = 1/[\mu_{m,x,y}(1 - \rho_{m,x,y}^{tp})] \\ \overline{D_{m,x,y}^{tr}} = 1/[\iota_{m,x,y}(1 - \rho_{m,x,y}^{tr})] \end{cases} \quad (33)$$

It is obvious that the constraint in Equation 34 should be met for rtPS/ErtPS connections to meet the target packet loss rate constraint.

$$\overline{D_{m,x,y}^{tp}} + \overline{D_{m,x,y}^{tr}} \leq D_{m,x,y} \quad (34)$$

4.3 BR performance analysis

We first analyze the BR delay saved by our enhanced BR scheme. Let $\tau_{m,x,y}$ be the unicast polling interval of the *BRU* of the uplink connection $C_{m,x,y}$. Since queuing delay in *tp* buffer is identical with the BR delay of our enhanced BR scheme, the average BR delay of an rtPS/ErtPS connection saved by our proposed BR scheme is

$$\mathrm{SD}_{m,x,y} = \tau_{m,x,y} - \overline{D_{m,x,y}^{tp}} \quad (35)$$

Using multicast/broadcast polling, each nrtPS/BE connection is taken as a unit to request bandwidth. Suppose SS requests bandwidth for an nrtPS/BE connection when Equation 19 is met, the BR time of an nrtPS/BE

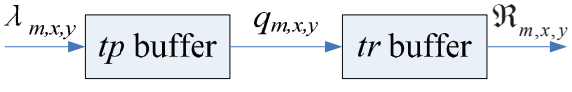

Figure 4 Queuing model for uplink transmission.

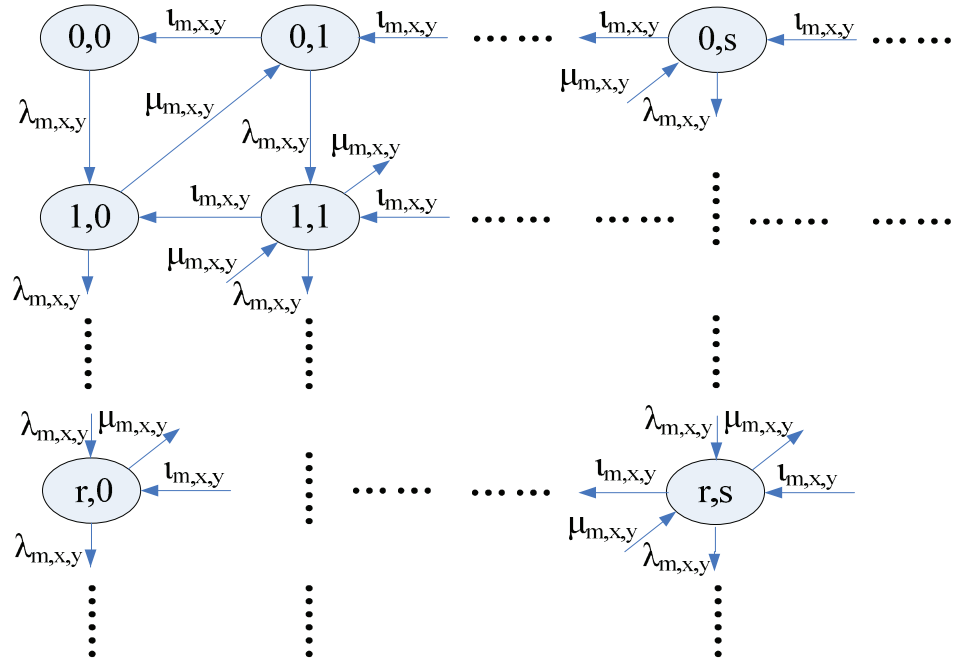

Figure 5 State transition diagram of two-dimensional Markov model.

connection also follows exponential distribution with the mean value of $1/\varsigma_{m,x,y}$, which is the collected effect of Equation 19 and BR retransmission in case of collision happening. F is the frame duration. In cell ν, the average number of contention BR messages transmitted in one frame can be deduced as

$$T_v = \left\lceil \sum_{m=1}^{M_v} \sum_{\gamma=1}^{Y_{i,3}} F\varsigma_{m,3,\gamma} + \sum_{m=1}^{M_v} \sum_{\gamma=1}^{Y_{i,4}} F\varsigma_{m,4,\gamma} \right\rceil \quad (36)$$

There are N BR opportunities for multicast/broadcast polling in one frame. The average collision probability is

$$\Delta_v = 1 - \binom{N}{1} * (N-1)^{T_v}/N^{T_v} = 1 - \left(1 - \frac{1}{N}\right)^{T_v - 1} \quad (37)$$

Before a contention BR message can successfully be received by BS, it may meet c times of collisions. The average collision time of a contention BR message is

$$E(c) = \sum_{c=0}^{\infty} c(1-\Delta_v)\Delta_v^c = (1-\Delta_v)\Delta_v \left(\frac{d}{d\Delta_v}\left(\sum_{c=0}^{\infty} \Delta_v^c\right)\right) = \frac{\Delta_v}{(1-\Delta_v)} \quad (38)$$

The time before a contention BR message can successfully be transmitted is deduced as

$$W_{m,x,\gamma} = \pi_{m,x,\gamma} E(c) = \Delta_v \pi_{m,x,\gamma}/(1-\Delta_v) \quad (39)$$

where $\pi_{m,x,y}$ is the retransmission interval of an nrtPS/BE connection $C_{m,x,y}$. Compared with multicast/broadcast polling, the BR delay saved by our scheme for nrtPS/BE connection is

$$\mathrm{SD}_{m,x,\gamma} = \frac{\delta_{m,x,\gamma}}{(1-\delta_{m,x,\gamma})\omega_{m,x,\gamma}} + F + W_{m,x,\gamma} - \overline{D^{tp}_{m,x,\gamma}} \quad (40)$$

where $\delta_{m,x,y} = \lambda_{m,x,y}/\omega_{m,x,y}$.

Considering the signaling overhead for BR, we first deduce the resource utilization of the contention period for multicast/broadcast polling as

$$RU = (1-\Delta_v) * T_v/N = (1-1/N)^{T_v - 1} * T_v/N \quad (41)$$

Figure 6 shows the numerical results of Equations 37 and 41 when T_v is 80. We find that small contention size cause low BR resource utilization because of the high collision probability. Even though larger contention size ensures lower collision probability, more contention BR opportunities are left unused, which in turn reduce the BR resource utilization as well. Since the highest BR resource utilization of the contention period equals 37%, we can conclude that high signaling overhead is caused by multicast/broadcast polling.

Using multicast/broadcast polling, the average ratio of successful BR transmission has a tradeoff with BR delay [27,28]. Thus, Equation 42 is defined to find out the optimal contention period size for multicast/broadcast polling. In Equation 42, larger θ_v indicates higher successful BR transmission rate and smaller average delay

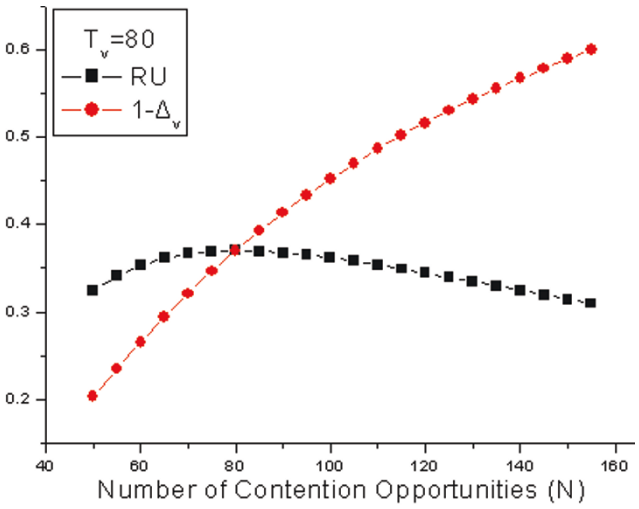

Figure 6 The relationship between BR success rate and the resource utilization of the contention period.

induced by collision. Therefore, the optimal contention period size is obtained when θ_v is maximized.

$$\theta_v = \left(\sum_{m=1}^{M_v}\sum_{\gamma=1}^{Y_{m,3}} \frac{1-\Delta_v}{W_{m,3,\gamma}} + \sum_{m=1}^{M_v}\sum_{\gamma=1}^{Y_{m,4}} \frac{1-\Delta_v}{W_{m,4,\gamma}}\right) \Big/ \left(/\sum_{m=1}^{M_v}(Y_{m,3}+Y_{m,4}\right) \quad (42)$$

Assuming $\pi_{m,x,y} = \pi$ is met for all uplink nrtPS/BE connections, Equation 42 can be simplified as

$$\theta_v = (1-\Delta_v) * \frac{(1-\Delta_v)}{\pi\Delta_v} \le \left((1-\Delta_v)^2 + \left(\frac{1-\Delta_v}{\pi\Delta_v}\right)^2\right) \Big/ 2 \quad (43)$$

The maximum value of θ_v is obtained when $\Delta_v = 1/\pi$. Using Equation 37, the optimal value of contention period size is deduced as

$$N_o = 1/\left(1 - \sqrt[T_v-1]{(\pi-1)/\pi}\right) \quad (44)$$

Let $\mathbb{M}$ denote the total number of nrtPS/BE *BRU*s which may request bandwidth in one frame duration using our proposed BR scheme. We have

$$\mathbb{M} = \left\lceil \sum_{m=1}^{M_v} F/\overline{D^{tp}_{m,3,-1}} + \sum_{m=1}^{M_v} F/\overline{D^{tp}_{m,4,-1}} \right\rceil \quad (45)$$

Therefore, compared with the optimal case of multicast/broadcast polling, the BR signaling overhead saved by our proposed BR scheme is N_o - $\mathbb{M}$.

5 Simulation results

Following assumptions apply in the simulation:

(a) There are 50 cells in our simulation environment. The BS in each cell communicates with the BSs in its neighbor cells to exchange handoff-related information, and the bandwidth reserved for the handoff connections is refreshed every 3 s. The symbol rate in each cell is 20 MBd, and the frame duration is 1 ms.

(b) In the initial status, there are 5,000 mobile SSs uniformly distributed over all cells. When an SS has intention to move from cell x to cell y, its handoff probability is a random value determined by its initial state. Under Rayleigh fading channel, all SSs have identical average SNR $\bar{\gamma}_m$, which equals 13.8.

(c) The new arriving connections are uniformly distributed in different mobile SSs. The probabilities of a new connection belonging to UGS, rtPS, ErtPS, nrtPS, and BE are 10, 25, 35, 20, and 10%, respectively, and, in unit of kb/s, the values of $[R^{\min}_{m,x,y} R^{\max}_{m,x,y}]$ for UGS, rtPS, ErtPS, nrtPS, and BE are [128, 128], [96, 386], [16, 64], [48, 128], and [0, 32], respectively.

(d) The intervals of unicast polling timers for rtPS/ErtPS *BRU*s equals 6 ms, while those for nrtPS and BE *BRU*s are 10 and 12 ms, respectively, and, the BR retransmission intervals of nrtPS and BE connections for multicast/broadcast polling are 6 and 8 ms, respectively.

Figure 7 shows comparison result of BR signaling overhead. As expected, our proposed enhanced BR scheme greatly reduces the signaling overhead compared with multicast/broadcast polling.

Using the proposed resource allocation and scheduling algorithm for packet-level QoS provision, our simulation compares the PA for AC and RR with the efficient AC scheme (EAC) [7] to evaluate the enhancement in CBR, CDR, CFR, and resource utilization with the growth of the connection arrival rate (CAR). Since the study of [7] assumes the system capacity is fixed, we consider three scenarios for performance comparison, which is, respectively, denoted as EAC1, EAC2, and EAC3. EAC1, EAC2, and EAC3 assume that the resource consumption for transmitting one bit in MAC layer equals $1/MR_1$, $1/$

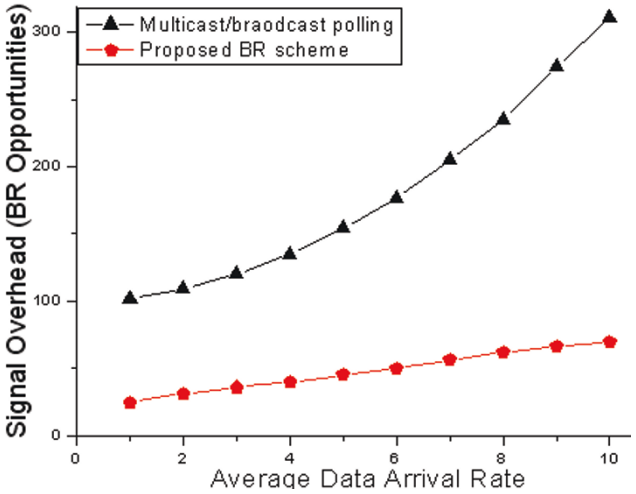

Figure 7 Signaling overhead comparison.

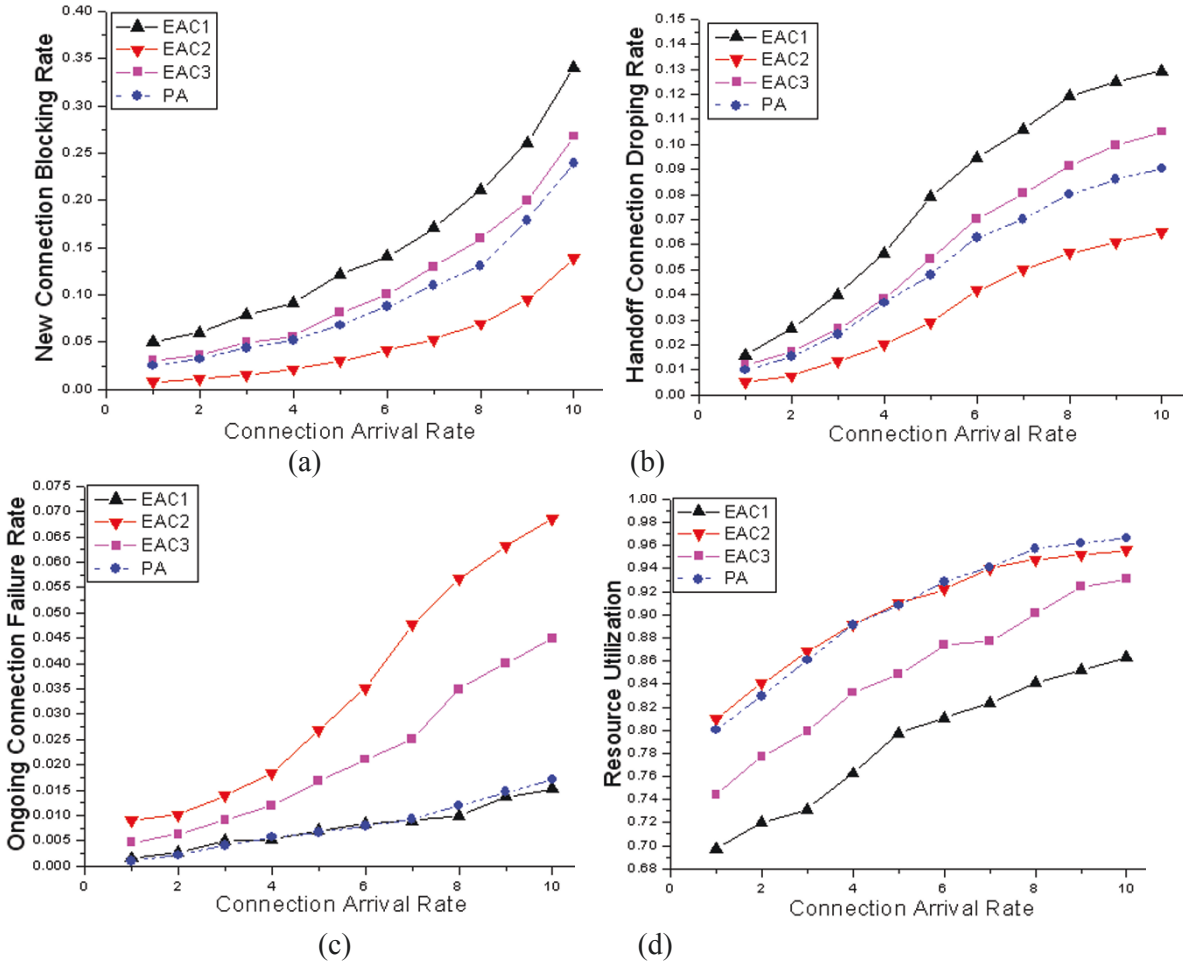

Figure 8 Performance evaluation under different CAR. **(a)** CBR versus CAR; **(b)** CDR versus CAR; **(c)** CFR versus CAR; **(d)** resource utilization versus CAR.

MR_K, and $\mathbb{C}^{avg}$, respectively. From the simulation results shown in Figure 8, we have the following conclusions:

• EAC1 undervalues the average system capacity, and the number of the connections accepted by the system is reduced accordingly. Therefore, EAC1 provides the best system stability with lowest CFR at the cost of highest CBR and CDR, as well as lowest resource utilization. However, our PA can achieve almost the same CFR as EAC1 while maintaining much better CBR, CDR, and resource utilization.

• More connections can be accepted in EAC2, which overvalues the system capacity. Therefore, CBR, CDR, and resource utilization are improved. However, the system becomes more unstable because of the high CFR. Even though the CBR and CDR of our approach are higher compared with EAC2, their resource utilizations almost keep the same. We can find that accepting more new/handoff connections at the cost of losing ongoing connections do not help to improve the resource efficiency.

• Our approach outperforms EAC2 in the four performance metrics. EAC2 only considers the average resource consumption for AC, while our approach considers one more restriction: the practical symbol consumption. If only for this reason, more new/handoff connections should be accepted in EAC2, and lower CBR and CDR are attained accordingly. However, compared with statistic RR in [7], our approach alleviates both resource over-reservation and insufficient reservation, which causes more new/handoff connections can be accepted.

From the theoretical analysis and the simulation results, it is noted that our proposed algorithm well balances the tradeoff among CBR, CDR, and CFR, while achieving high spectrum efficiency.

To evaluate the performance of packet-level QoS provision and system throughput using our enhanced BR scheme as well as joint resource allocation and scheduling algorithm, the scheduling algorithms in [20-22] are coupled with the traditional BR scheme of IEEE 802.16e to perform uplink transmission, which are denoted as approach1, approach2, and approach3, respectively. The performance metrics are evaluated with the growth of average data arrival rate (DAR). Particularly, the highest average DAR will cause the network enter into the saturated status, while the lowest average DAR can use up all resources under the lowest transmission rate (i.e., using MCS 1 for data transmission). From the simulation results illustrated in Figure 9, we have the following conclusions:

• The performance of packet-level QoS provision for rtPS/ErtPS is evaluated in terms of average PDR and maximum PDR variance. In Figure 9a,b, we find that our approach outperforms the others in the two metrics for real-time connections. The reason for ensuring lowest PDR is that our BR scheme reduces the BR delay, while our BA algorithm effectively avoids large real-time data being blocked under the deteriorated channel condition, which in turn reduces their scheduling delay. In addition, since fairness is ensured in case of bandwidth shortage to share the packet loss for the users with deteriorated channel conditions, smallest PDR variance is achieved. Since the connections with lower priority cannot use the bandwidth before all bandwidth requirements of high priority connections are satisfied in approach2, it works better than approach3, and approach1 produces highest PDR and largest packet dropping variance without QoS consideration.

• QoS outage rate is adopted to evaluate the QoS performance of nrtPS connections, which is the ratio of the "*QoS rate*" dissatisfaction times and the total BA times. Figure 9c shows that our approach provides the best QoS performance for nrtPS connections, because approach1 has no QoS consideration, approach2 causes the real-time connections preempting too much bandwidth for non-real-time connections, and approach3 does not consider QoS provision for nrtPS connections.

• Let G_m be the total uplink bandwidth granted to SS m in each frame. Based on resource allocation results in BS, from MAC viewpoint, the system throughput of cell ν can be deduced as Equation 46. It is obvious that the system throughput is a variable determined by the bandwidth requirement information of heterogeneous traffics, the QoS constraints, the channel conditions, and the resource allocation algorithm. Figure 9d illustrates the system throughput comparisons. When the bandwidth has not been used up under the low DAR, our approach achieves the highest throughput because the uplink bandwidth requirements can be reflected to BS more quickly. With the increase of the DAR, approach1 achieves the highest throughput because QoS provision compromise the system throughput gain of our approach, while approach2 provides the lowest throughput without considering the channel conditions. And, our approach works better than approach3 because we reduce the signaling overhead for BR.

$$Th_\nu = \sum_{m=1}^{M_\nu} G_m / MR_{k_m} \quad \text{s.t.} \quad \sum_{m=1}^{M_\nu} G_m \le TS_\nu \quad (46)$$

From the simulation results in Figure 9, it is noted that our PA well balances the tradeoff between packet-level QoS provision and spectrum efficiency.

6 Conclusions

Based on AMC in IEEE 802.16e PHY layer and flexible connection-oriented QoS provision in its MAC layer, this article investigates analytical integrated framework and adaptive QoS provision mechanism based on cross-

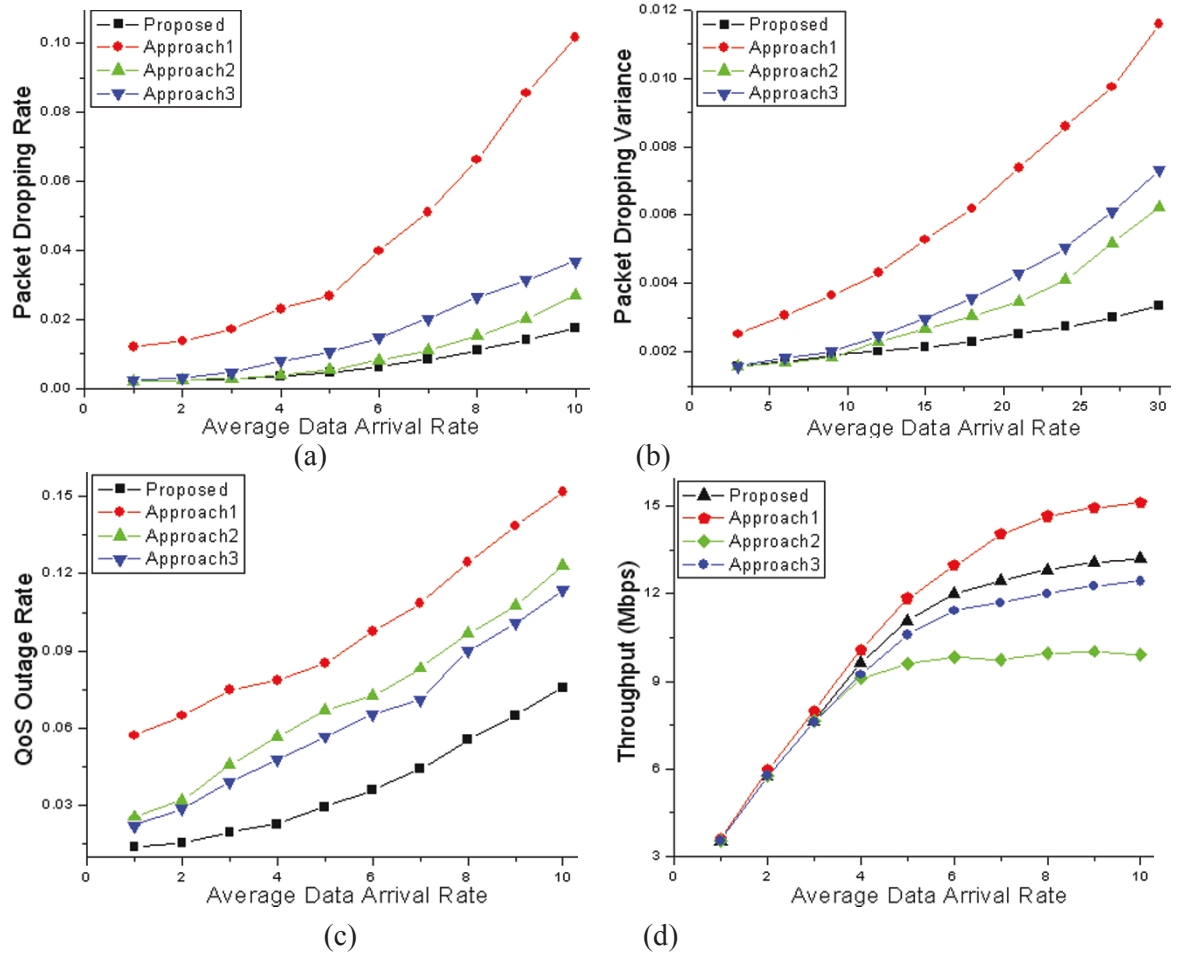

Figure 9 Connection-level QoS provision under different DAR: **(a)** PDR versus DAR; **(b)** PDR variance versus DAR; **(c)** QoS outage rate versus DAR; **(d)** system throughput versus DAR.

layer design. First, we propose an integrated framework for adaptive QoS provision cross-layer-based design. Our major QoS provision mechanism concerns are about connection-level QoS provision through dynamic RR and AC, as well as packet-level QoS provision through joint resource allocation and packet scheduling. Second, to alleviate the resource over-reservation and insufficient reservation for handoff connections, we estimate the average reserved resource over the unreliable wireless channel considering the handoff probability. In addition, we perform AC based on both average resource consumption and practical resource consumption. Particularly, adaptive QoS management is used to perform average source rate compression for accepting more new/handoff connections, as well as average source rate increasing for improving the service quality and resource utilization. Finally, a joint resource allocation and packet scheduling algorithm is designed to provide packet-level QoS guarantee in term of "*QoS rate*", which provides fairness for the services with identical scheduling priority in case of bandwidth shortage. In addition, we enhance the BR mechanism to reduce the BR delay and signaling overhead, which belongs to packet-level QoS provision. The theoretical analyses and the simulation results show that our approach guarantees well the QoS requirements of heterogeneous services, as well as provides high spectrum efficiency.

Acknowledgements

This study was supported by the Fundamental Research Funds for the Central Universities (2011RC0112), NSFC (60972076, 61072052), and Important National Science & Technology Specific Projects (2010ZX03003-004-03).

Author details

[1]Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, PR China [2]National Chengchi University, Taipei, Taiwan [3]SAP Research CEC, Zurich, Switzerland

Competing interests
The authors declare that they have no competing interests.

Received: 31 January 2011 Accepted: 19 August 2011
Published: 19 August 2011

References

1. IEEE 802.16 Standard-Local and Metropolitan Area Networks-Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE 802.16-2004
2. IEEE 802.16 Standard-Local and Metropolitan Area Networks-Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. IEEE 802.16e-2005
3. Q Liu, S Zhou, GB Giannakis, Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks, IEEE J Sel Areas Commun. **23**(5), 1056–1066 (2005)
4. J Ye, J Hou, S Papavassiliou, A comprehensive resource management framework for next generation wireless networks, IEEE Trans Mobile Comput. **1**(4), 249–264 (2002). doi:10.1109/TMC.2002.1175539
5. H Wang, W Li, DP Agrawal, Dynamic admission control and QoS for 802.16 wireless MAN, in *Proc of Wireless Telecomm Symp*, 60–66 (April 2005)
6. D Niyato, E Hossain, Joint bandwidth allocation and connection admission control for polling services in IEEE 802.16 broadband wireless networks, in *Proc of IEEE Int Conf on Commun*. **12**, 5540–5545 (June 2006)
7. H Yao, GS Kuo, A QoS-Adaptive admission control for IEEE 802.16e-based mobile BWA networks, in *Proc of IEEE Consumer Comm and Networking Conf*., 833–837 (January 2007)
8. Y Ge, GS Kuo, An efficient admission control scheme for adaptive multimedia services in IEEE 802.16e networks, in *Proc of IEEE Veh Tech Conf*., 1–5 (September 2006)
9. K Gakhar, M Achir, A Gravey, Dynamic resource reservation in IEEE 802.16 broadband wireless networks, in *the 14th IEEE Int Workshop on Quality of Service*, 140–148 (June 2006)
10. X Guo, W Ma, Z Guo, Z Hou, Dynamic bandwidth reservation admission control scheme for the IEEE 802.16e broadband wireless access systems, in *Proc of IEEE Wireless Comm and Networking Conf*., 3418–3423 (January 2007)
11. DA Levine, IF Akyildiz, M Naghshineh, A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept, IEEE/ACM Trans Network. **5**(1), 1–12 (1997). doi:10.1109/90.554717
12. H Chen, S Kumar, CCJ Kuo, Dynamic call admission control scheme for QoS priority handoff in multimedia cellular systems, Proc IEEE Wirel Commun Network. **1**, 114–118 (2002)
13. MH Ahmed, Call admission control in wireless networks: a comprehensive survey, IEEE Commun Surv Tutor. **7**(1), 50–69 (2005)
14. D Niyato, E Hossain, Call admission control for QoS provisioning in 4G wireless networks: issues and approaches, IEEE Network. **19**(5), 5–11 (2005). doi:10.1109/MNET.2005.1509946
15. N Nasser, H Hassanein, Prioritized multi-class adaptive framework for multimedia wireless networks, in, *Proc IEEE Int Conf Commun* **7**, 4295–4300 (2004)
16. L Huang, S Kumar, C-CJ Kuo, Adaptive resource allocation for multimedia QoS management in wireless networks, IEEE Trans Wirel Technol. **53**(2), 547–558 (2004)
17. M Wang, GS Kuo, A QoS-adaptive resource reservation scheme for MPEG4-based services in wireless networks, in *Proc IEEE Int Conf Commun*. **5**, 16–20 (May 2005)
18. G Schembra, A resource management strategy for multimedia adaptive-rate traffic in a wireless network with TDMA access, IEEE Trans Wirel Commun. **4**(1), 65–78 (2005)
19. E Kwon, J Lee, K Jung, S Ryu, A performance model for admission control in IEEE 802.16, in Lecture Notes in Computer Science, vol. 3510 (Springer, 2005), pp. 159–168. doi:10.1007/11424505_16
20. R Knopp, PA Humblet, Information capacity and power control in single cell multiuser communications, in *Proc of IEEE Int Conf on Commun*, 331–335 (June 1995)
21. K Wongthavarawat, A Ganz, IEEE 802.16 based last mile broadband wireless military networks with quality of service support, *Proc of IEEE Military Comm Conf*. **2**, 779–784 (October 2003)
22. SJ Sang, DG Jeong, WS Jeon, Cross-layer design of packet scheduling and resource allocation in OFDMA wireless multimedia networks, in *Proc of IEEE Veh Tech Conf*., **1**, 309–313 (September 2006)
23. F Hou, PH Ho, X Shen, AY Chen, A novel QoS scheduling scheme in IEEE 802.16 networks", in *Proc of IEEE Wireless Comm and Network Conf*., 2457–2462 (March 2007)
24. M Andrews, K Kumaran, K Ramanan, A Stolyar, P Whiting, R Vijayakumar, Providing quality of service over a shared wireless link, IEEE Commun Mag. **39**(2), 150–154 (2001). doi:10.1109/35.900644
25. Q Liu, X Wang, GB Giannakis, A cross-layer scheduling algorithm with QoS support in wireless networks, IEEE Trans Veh Technol. **55**(3), 39–847 (2006)
26. H Lee, DH Cho, An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system, IEEE Commun Lett. **9**(8), 691–693 (2005). doi:10.1109/LCOMM.2005.1496584
27. SM Oh, JH Kim, The analysis of the optimal contention period for broadband wireless access network, in *Proc of Int Conf on Pervasive Computing and Commun Workshops*., 215–219 (March 2005)
28. J Yan, GS Kuo, Cross-layer design of optimal contention period for IEEE 802.16 BWA systems, in *Proc of IEEE Int Conf on Commun*. **4**, 1807–1812 (June 2006)
29. A Vinel, Y Zhang, M Lott, A Tiurlikov, Performance analysis of the random access in IEEE 802.16, in *Proc of IEEE Int Symp on Personal, Indoor and Mobile Radio Commun*. **3**, 1596–1600 (September 2005)
30. R Iyengar, V Sharma, K Kar, B Sikdar, Analysis of contention-based multi-channel wireless MAC for point-to-multipoint networks, in *Proc of Int Symp on a World of Wireless, Mobile and Multimedia Networks*., 1–3 (June 2006)
31. L Lin, W Jia, W Lu, Performance analysis of IEEE 802.16 multicast and broadcast polling based bandwidth request", in *Proc of IEEE Wireless Comm and Network Conf*., 1854–1859 (March 2007)
32. SM Oh, JH Kim, The optimization of the collision resolution algorithm for broadband wireless access network, in *Proc of Int Conf on Advanced Comm Technol*. **3**, 1944–1948 (February 2006)
33. S Shakkottai, TS Rappaport, PC Karlsson, Cross-layer design for wireless networks, IEEE Commun Mag. **41**(10), 74–80 (2003). doi:10.1109/MCOM.2003.1235598
34. RA Berry, EM Reb, Cross-layer wireless resource allocation, IEEE Signal Process Mag. **21**(5), 59–68 (2004). doi:10.1109/MSP.2004.1328089
35. GS Kuo, PC Ko, ML Kuo, A Probabilistic resource estimation and semi-reservation scheme for flow-oriented multimedia wireless networks, IEEE Commun Mag. **39**(2), 135–141 (2001). doi:10.1109/35.900642
36. EL Hahne, Round-Robin scheduling for max-min fairness in data networks, IEEE J Sel Areas Commun. **9**(7), 1024–1039 (1991). doi:10.1109/49.103550

doi:10.1186/1687-1499-2011-69
Cite this article as: Zhang *et al.*: Adaptive QoS provision for IEEE 802.16e BWA networks based on cross-layer design. *EURASIP Journal on Wireless Communications and Networking* 2011 **2011**:69.