A Capacity-Optimal QoS Provisioning Scheme for Multimedia Traffic in CDMA Networks

Tao Shu and Zhisheng Niu
State Key Lab on Microwave and Digital Communication
Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China

Abstract — We study the transmission rate ($R$) and bit-energy-to-interference ratio ($E_b/I_0$) needed to maintain the desired QoS requirement for a given traffic while maximizing the user capacity in multimedia CDMA networks. Closed form functions among transmission rate, bit error rate (or equivalently $E_b/I_0$), QoS requirements, and traffic characteristics are derived. Based on the dependence between $R$ and $E_b/I_0$, we propose a capacity-optimal determination of ($R$, $E_b/I_0$) for a given traffic and QoS requirement. By numerical example we show that higher user capacity is achieved by our scheme than those which ignore the dependence between $R$ and $E_b/I_0$.

Key words — QoS provisioning, CDMA, capacity optimization

I. INTRODUCTION

The next generation wireless networks will be expected to provide respective quality-of-service (QoS) support for heterogeneous traffic having different characteristics and QoS requirements. Because of the advantage in user capacity, the Code Division Multiple Access (CDMA) is a competitive multiple access technique in the realization of next generation wireless networks. In the CDMA system, the data transmission power and the transmission rate are two elements which must be determined to provide QoS guarantees for the traffic. To date, extensive research has been conducted on the determination of these two elements[1,2]. In these papers, the transmission power and rate are computed optimally to minimize the total transmission power or to maximize the total transmission rate among all users in a cell. Users belonging to the same traffic class are associated with the same transmission rate and bit-energy-to-interference ratio requirement ($R$, $E_b/I_0$). However, the determination of $E_b/I_0$ and $R$ for each traffic class is ignored in most current research. In this paper, we show that by appropriate determination of $R$ and $E_b/I_0$, we can not only satisfy the user’s QoS requirement but also optimize user capacity for the whole system. Our idea includes two parts: firstly we derive the set of ($R$, $E_b/I_0$) which satisfy user’s QoS requirement by fluid approximation analysis and then we propose a capacity-optimal scheme to select the most appropriate ($R$, $E_b/I_0$) value among this set which maximizes the number of users that the system can admit. The link layer error recovery scheme of hybrid-ARQ is also included in the analysis.

Several studies have been conducted on the relation between QoS and transmission bandwidth (or effective bandwidth) in wireless networks. In [3], the authors proposed the effective bandwidth as a metric of quality of service for wired and wireless ATM. However, the impact of error control was not considered. In [4] and [5], the authors proposed the effective bandwidth subject to packet loss and subject to delay bound, respectively. All these studies were conducted under general wireless environment, only the effective bandwidth was concerned and $E_b/I_0$ was ignored. Different from the previous works, we concern the trade-off relation between $R$ and $E_b/I_0$ satisfying user’s QoS and present its importance in the user capacity optimization in CDMA networks.

The rest of this paper is organized as follows. In section II, we describe the wireless link model. In section III, we derive the relation among $R$, $E_b/I_0$, and QoS requirements for real time and non-real time users, respectively. In section IV, we propose the capacity-optimal determination of $R$ and $E_b/I_0$ for a given user. Numerical example and discussions are reported in section V and we conclude our work in section VI.

II. WIRELESS LINK MODEL

We consider a CDMA system in Fig.1. We take the downlink (from base to mobiles) as an example, where multiple transmissions happen simultaneously at their individual power levels and transmission rates. Each wireless link shares a fraction of the total capacity. There are two types of queues, one for the non-real time traffic and the other for the real time traffic. The non-real time traffic can be the traffic of web browsing or FTP and the real time traffic can be such traffic as voice or interactive applications on the web. Packets generated to different mobile stations are stored in different queues. Thus we can study the queue separately. A wireless link can be equivalent to the queuing model shown in Fig.2.

For the non-real time traffic, we assume the concerned QoS is the packet loss probability $P_{over}$. The finite
buffer (buffer size is $K$ packets) overflow. For the real time traffic, we assume the concerned QoS is the packet discarding rate $P_{\text{delay}}$ due to delay bound ($T_{\text{bound}}$) excess, i.e., $Pr\{\text{delay}>T_{\text{bound}}\}<P_{\text{delay}}$. The packets in the queue are served according to a FIFO rule. A hybrid-ARQ link-layer recovery scheme is employed. After the CRC and FEC coding, the correctable bit number of a packet is $r$ and the packet length is $L$ bits. We assume a stop-and-go scheme is employed in the transmission, i.e., the transmitter will stop to wait for the ACK/NACK after it transmits one packet. The transmission rate is $R$ packets/s and the transmission time of a packet is $T=1/R$ s. After the received packet is error decoded, if there is still error in the packet, the receiver will send a NACK to the transmitter, or an ACK if all the errors can be corrected. We ignore the coding/decoding time and the transmission time of ACK/NACK. If a NACK is received by the transmitter, it will retransmit the head of queue immediately until it receives an ACK to the packet. And then it starts to transmit the next packet. We assume there is no limitation of the largest retransmission number in the hybrid-ARQ. A time-stamp checker is used in the real time traffic queue in Fig.1 to check whether the delay bound is exceeded and the packets exceeding the delay bound is discarded.

![Fig.1 Wireless link model for a multimedia CDMA system.](image)

We assume a perfect power control for the transmission, i.e., the received $E_b/I_0$ will be maintained at its target. Considering the multiple access interference (MAI) on the channel, let the BER for the transmission be $P_{\text{be}}$ which is determined uniquely by received $E_b/I_0$. The relation between $P_{\text{be}}$ and $E_b/I_0$ is determined by the modulation/demodulation and coding/decoding scheme in the physical layer and is generally assumed as a monotonic decreasing function:

$$P_{\text{be}} = \Phi(\frac{E_b}{I_0})$$  \hspace{1cm} (1)

Because of the one-to-one relation between $P_{\text{be}}$ and $E_b/I_0$, we use these two terms interchangeably in this paper. The probability of a successful packet transmission is given by

$$P_s = \sum_{i=0}^{r} \frac{(L^i)}{i!} P_{\text{be}} ^i (1-P_{\text{be}}) ^{L-i}$$  \hspace{1cm} (2)

thus the probability of an unsuccessful packet transmission is simply given by $P_c = 1 - P_s$.

The transmission number $N$ used to deliver a packet successfully (including the first transmission and the retransmission followed) is a random variable which follows the geometric distribution with the parameter $P_c$:

$$Pr\{N = j\} = P_c ^{j-1} P_s , \quad j = 2, 3, \ldots$$  \hspace{1cm} (3)

and the average value of $N$ is: $\overline{N} = 1/P_s$. The time needed to transmit a packet successfully is $D = NT = \frac{N}{R}$ and its average value is $\overline{D} = \frac{1}{PR}$.

Thus the average service rate for a packet is $R = PR$.

### III. FLUID ANALYSIS FOR QoS PROVISIONING

An ON-OFF source with peak packet rate $g$ is assumed to model the bursty nature of the traffic. The ON and OFF periods are exponentially distributed with mean $1/\mu$ and $1/\lambda$, respectively.
A. Non-real time traffic case

The assumed QoS requirement for the non-real time traffic is the packet loss probability due to the buffer overflow.

Let \( P(x) \) (\( i=0 \text{ or } I \), where 0 represents OFF state and 1 represents ON state) be the cumulative density function (CDF) of the queue length \( l \) when the system is in state \( i \). Let \( P(x) = [P_0(x), P_1(x)] \). Following a standard fluid approach, we have

\[
\frac{dP(x)}{dx}D = P(x)M
\]

where \( D \) is the drift matrix and \( M \) is the infinitesimal generator of the Markov process representing the ON-OFF source, given by

\[
D = \begin{bmatrix} -R & 0 \\ 0 & g - R \end{bmatrix}, \quad M = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}
\]

The solution to Eq. (4) has the following form:

\[
P(x) = k_1 F_1 e^{\alpha_1 x} + k_2 F_2 e^{\alpha_2 x}
\]

where \( \alpha_1 \) and \( \alpha_2 \) are the eigenvalues of the matrix \( MD^{-1} \), the row vector \( F_1 \) and \( F_2 \) are the eigenvectors corresponding to \( \alpha_1 \) and \( \alpha_2 \), respectively, and the \( k_1 \) and \( k_2 \) are coefficients which can be determined by the boundary conditions.

By simple mathematical manipulation, we get the solutions to Eq. (4)

\[
P_0(x) = \frac{\mu}{\lambda + \mu} - \frac{\lambda (g - R)}{(\lambda + \mu)R} e^{\frac{\lambda}{\mu} \frac{R}{(g - R)}}
\]

\[
P_1(x) = \frac{\mu}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{\frac{\lambda}{\mu} \frac{R}{(g - R)}}
\]

Hence the CDF of the queue length is simply as follows

\[
P(x) = P_0(x) + P_1(x) = 1 - \frac{\lambda g}{(\lambda + \mu) R} e^{\frac{\lambda}{\mu} \frac{R}{(g - R)}}
\]

We should note that the \( x \) in the CDF above is a continuous variable, which represents the queue length in the fluid environment. The discretization of the queue length can be given by

\[
Pr(l = 0) = P(0) = \frac{\lambda g}{(\lambda + \mu) R}
\]

\[
Pr(l = k) = P(k) - P(k - 1) = \frac{\lambda g}{(\lambda + \mu) R} (e^{a} - 1) e^{-ak}
\]

where \( a = \frac{g - R}{R P_c} \). It is obvious that the queue length has a stable distribution only when \( a > 0 \), which means

\[
R > \frac{\lambda g}{\lambda + \mu} \quad \text{or} \quad R > \frac{\lambda g}{(\lambda + \mu) P_c}
\]

For a buffer with size \( K \), by summing up the probability of queue length from \( K+1 \) to infinity, we get the packet loss probability due to buffer overflow:

\[
P_{\text{overflow}} = \frac{\lambda g}{(\lambda + \mu) R P_c} e^{-ak}
\]

B. Real time traffic case

We assume the interested QoS requirement for real time traffic is the packet discarding probability due to delay bound violation, i.e., \( Pr\{\text{delay} > T_{\text{bound}}\} < P_{\text{delay}} \).

For an arbitrary packet, the delay time \( S \) it experiences is the sum of the service time of the packets in front of it (including the one in the transmission) on its arrival plus its own service time. Because the transmission number \( N \) for a packet follows the geometric distribution with parameter \( P_c \), the left service number for the packet in transmission also follows the geometric distribution because of the memorryless feature of geometric distribution. Assume there are \( k-I \) packets in front of it on the arrival of the incoming packet, and let \( N_i \) be the transmission number for the \( i \)th (\( i=1,2,..,k \)) packet, then the delay time \( S \) should be:

\[
S = \sum_{j=1}^{k} N_j T = N_{\text{total}} T
\]

where the compound random variable \( N_{\text{total}} \) is the total transmission number between the arrival and the successful departure of the packet referred.

The probability that there are \( k-I \) packets in front of the incoming packet on its arrival can be approximated by the probability that given the source is in ON state there are \( k \) packets in the system. From Eq. (7), we get:

\[
P_{\text{on}}(l < x) = Pr\{l < x | ON\} = \frac{P(x)}{\lambda} = 1 - e^{-ax}
\]

By discretization of Eq. (13), the probability that given the source is in ON state there are \( k \) packets in the system is given by:

\[
P_{\text{on}}(l = k) = P_{\text{on}}(l \leq k) = P_{\text{on}}(l \leq k - 1)
\]

\[
= (e^a - 1) e^{-ak}, \quad k = 1,2,3...\]

By applying embedded generating function method, the generating function \( H(z) \) of the compound random variable \( N_{\text{total}} \) can be given by

\[
H(z) = (e^{a} - 1) \frac{e^{-ax} P_c z}{1 - (P_c + e^{-ax} P_c) z}
\]

\[
= (e^a - 1) e^{-ax} P_c \sum_{j=1}^{\infty} (P_c + e^{-ax} P_c)^{-1} z^j
\]

which indicates the probability that \( N_{\text{total}} \) is equal to \( j \) should be:

\[
Pr[N_{\text{total}} = j] = (e^{-a}) e^{-ax} P_c (P_c + e^{-ax} P_c)^{-1} j \geq 1,2.
\]

For a given delay bound \( T_{\text{bound}} \), let \( n \) be the maximum integer which is not larger than the \( T_{\text{bound}}/T \), then the packet loss probability due to delay bound violation is approximated by:

\[
P_{\text{delay}} = Pr\{S > T_{\text{bound}}\} \leq Pr\{N_{\text{total}} \geq n\} = \sum_{j=n}^{\infty} Pr[N_{\text{total}} = j]
\]

\[
= (P_c + e^{-ax} P_c)^{-1}
\]
C. The trade-off between R and E_b/I_0

Both Eq.(11) and Eq.(17) show that more than one pair of (R, P_{be}) can satisfy the specified QoS requirement for a given traffic. This indicates a trade-off relation between R and P_{be}. Figs.3-4 plot the (R, P_{be})s which satisfy the QoS requirements for non-real time and real time users according to Eq.(11) and (17), respectively. The parameters of the assumed ON-OFF source is g=64kb/s, \lambda = -6.01, \mu = -4.01, and the length of each packet is 512 bits.

In Fig.3 and Fig.4, the points on the same curve consist the set of (R, P_{be})s satisfying the same QoS requirement. This results in an optimization issue: a higher BER will require a lower Eb/I0, which reduces the transmission power needed and increases user capacity; at the same time, in order to achieve the desired QoS a larger transmission rate is needed which reduces the processing gain and thus reduces the user capacity. There should be a balancing point of (R, P_{be}) by which maximum user capacity is achieved. In section IV we present our capacity-optimal scheme for the determination of the appropriate (R, P_{be}).

IV. CAPACITY-OPTIMAL DETERMINATION OF (R, P_{be})

We only consider a single cell CDMA system. Assume that there are M users in the system. The desired transmission rate of user i (i=1,2,...,M) is R_i and its received E_b/I_0 target is \gamma_i. The total system bandwidth is WHz. We assume the additive white Gaussian noise (AWGN) can be ignored compared to the MAI. As proved by [2], the system exists a feasible power and rate allocation to maintain each user acquiring its desired transmission rate and Eb/I0 target if and only if

$$g_i = \sum_{j=1}^{M} g_j < 1$$  \hspace{1cm} (18)

where

$$g_i = \frac{1}{W} \frac{1}{R_i \gamma_i}$$  \hspace{1cm} (19)

Eq.(18) indicates that the less the g_i is, the more users the system can admit. In order to maximize the user capacity of the system, the g_i of individual user should be minimized. This is equivalent to the minimization of the product of R_i and \gamma_i.

Our objective is to minimize the product of R_i and \gamma_i while promising the QoS requirement of the user can still be met: for non-real time traffic:

$$\min R_i \gamma_i \quad \text{s.t.} \quad P_{\text{overflow}} = \frac{\lambda g_i}{(\lambda + \mu) R_i P_i} e^{-\lambda R_i}$$  \hspace{1cm} (20)

and for real time traffic:

$$\min R_i \gamma_i \quad \text{s.t.} \quad P_{\text{delay}} \leq (P_{\text{ci}} + e^{-\mu} P_{\text{ci}})^{\gamma_i}$$  \hspace{1cm} (21)

where

$$P = \sum_{j=1}^{M} L_{ij} P_{ji} (1 - P_{ji})^{\gamma_j}$$ and $$P_{be} = \Phi(\gamma)$$, as defined by Eq.(2) and (1), respectively.

The optimal solution to Eq.(20) and (21) is the appropriate (R, E_b/I_0) which maximizes the system user capacity while satisfying user’s QoS requirement.

V. NUMERICAL EXAMPLE

Consider a voice/data WCDMA system with total bandwidth of 5MHz. For the voice traffic: the average active period is 0.375s, the average silent period is 0.625s, the peak bit rate during the active period is 16kb/s, and the packet length is 256 bits. The QoS requires that the percentage of packets whose delay exceeds 50ms should be less than 5%. For the data traffic: the average ON period is 1.714s, the average OFF period is 15.429s, the peak bit rate during ON
period is 56kb/s, and the data packet length is 512 bits. The average ON period is equivalent to the time it takes to transfer a 12K bytes web page (the average length of WWW pages [6]) at the rate of 56kb/s. The buffer size of each data user is 50k bits. The QoS requires that the packet loss rate due to buffer overflow should be less than $10^{-3}$. We assume the relation between $P_{be}$ and $E_b/I_0$ is:

$$P_{be} = \Phi\left(\frac{E_b}{I_0}\right) = \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{2} \left(\frac{E_b}{I_0}\right)^2}$$

We assign $(P_{be}=10^{-3}, R=15.8kb/s)$ to voice user for its QoS requirement. For the data user, we compare two QoS provisioning schemes. In the first scheme, we simply assign $(P_{be}=10^{-6}, R=49.77kb/s)$ for data user. We will call this scheme the traditional QoS provisioning scheme. In the second scheme, we use Eq.(20) for QoS provisioning. We call it the capacity-optimal QoS provisioning scheme. As shown in Fig.5, $(P_{be}=6\times10^{-4}, R=51.76kb/s)$ is selected as the optimal solution to Eq.(20). The system user capacities resulted by these two schemes are compared in Fig.6. Considerable improvement in system user capacity is achieved by the proposed capacity-optimal QoS provisioning scheme.

VI. CONCLUSION

We study the transmission rate ($R$) and bit-energy-to-interference ratio ($E_b/I_0$) needed to maintain the desired QoS requirement for a given traffic while achieving the maximum capacity in multimedia CDMA networks. Both real time and non-real time traffic are considered and link layer error recovery scheme hybrid-ARQ is included in the analysis. By applying fluid approximation method, closed form functions among transmission rate, bit error rate (or equivalently $E_b/I_0$), QoS requirements, and traffic characteristics are derived. The results indicate a trade-off relation between $R$ and $E_b/I_0$, satisfying the user’s QoS requirement. Based on the dependence between $R$ and $E_b/I_0$, we propose a capacity-optimal determination of $(R, E_b/I_0)$ for a given traffic and QoS requirement. By numerical example we verify that higher user capacity is achieved by our scheme than those which ignore the dependence between $R$ and $E_b/I_0$.

REFERENCES


