Journal of the Korean Institute of Industrial Engineers
Vol. 33, No. 1, pp. 153-163, March 2007.

Quality of Service Parameters Estimation Model for Adaptive Bandwidth Service in Mobile Cellular Networks

Sung Hwan Jung[1] · Jung-Wan Hong[2†] · Chang Hoon Lie[1]

[1] Department of Industrial Engineering, Seoul National University
[2] Department of Industrial Systems Engineering, Hansung University

적응형 서비스를 제공하는 이동통신망에서의 서비스 품질 척도 추정 모델

정성환[1] · 홍정완[2] · 이창훈[1]

[1] 서울대학교 산업공학과 / [2] 한성대학교 산업시스템공학과

An adaptive framework paradigm where the bandwidth values of the ongoing calls vary according to the traffic situations is one of the promising concepts for overcoming poor resource conditions due to handoffs in mobile cellular networks. However, quantifying the level of bandwidth degradation of the ongoing calls in an adaptive framework is important in view of Quality of Service (QoS) provisioning. Therefore we introduce new QoS parameters, the Degradation Degree Ratio (DDR), which represents the average portion of the degradation degree during degradation period of a call, and the Degradation Area Ratio (DAR), which represents the average ratio of a call's degradation level considering both the period and the degree of degradation jointly in multi-level bandwidth service. We also develop a new analytical model for estimating the QoS measures such as the Degradation Period Ratio (DPR), DDR and DAR. We show how to calculate the QoS measures and illustrate the method by numerical examples. The proposed model can be used to determine the optimal parameter of the CAC scheme and analyze the sensitivity of the QoS parameters in adaptive networks.

Keywords: Adaptive Framework, Degradation Period Ratio, Degradation Degree Ratio, Degradation Area Ratio, Bandwidth Adaptation Algorithm

1. Introduction

Quality of Service (QoS) guaranteeing in mobile cellular networks is becoming more and more important together with the increase in the demand on wireless mobile communications. The most significant QoS parameters in mobile cellular networks are new call blocking probability (NBP) and forced termination probability. The former is the probability that a new arrival call will be blocked and the latter is the probability that an ongoing call will be forced to terminate before the completion of service. It is generally accepted that forced termination of an ongoing call is much more unbearable to users than the blocking of a new call from the QoS aspect. In cellular networks, forced termination probability of calls is almost directly proportional to the handoff dropping probability (HDP), the

This research was financially supported by Hansung University in the year of 2006.
† Corresponding author : Professor Jung-Wan Hong, Department of Industrial Systems Engineering, Hansung University, Samseon-dong 3-ga, Seongbuk-gu Seoul 136-792, Korea, Tel : +82-2-760-4392, Fax : +82-2-760-4490, E-mail : jwhong@hansung.ac.kr
Received May 2006; revision received September 2006; accepted October 2006.

probability that handoff attempt fails. Here, handoff is the process of changing communicating base stations while an user crosses over the cell boundary. So far, call admission control (CAC) of cellular networks has focused on how to block new calls to reduce the HDP while maximizing the utilization of system bandwidth.

Recently, the cell size of cellular networks tends to be smaller due to the limited radio bandwidth and larger population of the wireless/mobile users. As a result, more handoff attempts happen through the life time of a call. This adds on difficulties to providing stable QoS levels in cellular networks. However, with the introduction of adaptive frameworks, a stable QoS provisioning is expected to become possible and HDP can be reduced to a negligible level in normal traffic load (Naghshineh and Willebeek-LeMair, 1997 : Bharghavan *et al*., 1998). In adaptive framework networks, the bandwidth values of ongoing calls can be dynamically adjusted according to the traffic situations.

The core technology, which makes an adaptive multimedia networking possible, is the adaptive multimedia encoding. The multimedia stream is compressed in the form of a layered or a hierarchical coding to support heterogeneous receivers. Therefore, each receiver or sender can selectively choose a subset of layered coding depending on both its capability and bandwidth availability. The multiple channel assignment is also a fundamental technology to give a body to an adaptive framework. Adaptive Multi Rate (AMR), which is the specified speech codec in WCDMA, is an example of an adaptive service, which can adjust its source rate from 12.2 kbps to 4.75 kbps. The widely used video codec standard, H.263 and MPEG4, are also adaptive (Zhao and Zhang, 2001).

When there are insufficient channels to accept incoming handoff calls, a Bandwidth Adaptation Algorithm (BAA) can be utilized to reduce the size of the bandwidth of the ongoing calls. Such a process of reducing is referred to as "bandwidth degradation". HDP which is an important QoS parameter in non-adaptive networks can be reduced to near zero if the minimum possible bandwidth of a call is sufficiently small. As a result, the forced termination of a handoff call does not need to be considered in adaptive networks any longer. Instead, new QoS parameters need to be proposed in order to quantify the level of bandwidth degradation of a call in view of QoS provisioning.

1.1 Related works

In the existing non-adaptive framework, most studies have been performed to define the optimal CAC policies for maximizing the resource utilization (minimizing call blocking) subject to a user's HDP constraint (Jain and Knightly, 1999). In adaptive mobile cellular networks, one of the existing studies is to determine an optimal BAA that considers diverse objectives such as maximizing revenue, minimizing cost and accomplishing fairness by using optimization techniques (Ahn and Kim, 2003 : Kwon *et al*., 1999).

Other existing studies are concerned with proposing new QoS measures for characterizing the level of degradation properly and designing CAC policies guaranteeing proposed QoS measures. To capture the level of bandwidth degradation in an adaptive framework, the Degradation Period Ratio (DPR) is proposed as a new QoS parameter by Kwon *et al*. (1998). This parameter represents the average portion of a call's life time when it is degraded. However, DPR does not characterize the degree of bandwidth degradation. Therefore Xiao *et al*. (2002) propose two QoS parameters, the degradation ratio (DR) and the degradation degree (DD) for multiple classes of adaptive multimedia services. These two QoS parameters characterize both the frequency of degradation as well as the degree of degradation. However, they are not presented as analytical formulas but as measurement-based ones using a time averaging method. Kwon *et al*. (2003) propose a distributed CAC algorithm that guarantees the upper bound of the cell overload probability. They also present a BAA, which seeks to minimize the cell overload probability. Nasser and Hassanein (2004) derived the QoS metrics such as NBP, HDP and DP using multi-dimensional Markov chain under multi-classes situation.

Other existing QoS measures are the Degradation Probability (DP) (Kwon *et al*., 1999), the Degradation Area Size (Xiao *et al*., 2000) and the utility function (Wang *et al*., 2001), etc. Most of these parameters give significance to system providers. However, they do not give definite meaning to service users as QoS measures. Chou and Shin (2004) propose two new user-perceived QoS metrics, degradation ratio and upgrade/degrade frequency in multi-level degradation. They also provide the multi-dimensional Markov model to derive these two QoS metrics in single-class case. However, the complexity of model is too high to be applied to multi-class case and the way to find optimal

control parameter is not presented.

1.2 Our contributions

We develop a new analytical model with low complexity for estimating QoS parameters more realistically, which provides definite meaning from the service user's perspective in adaptive framework networks. In most of previous analytical studies, the QoS parameters are defined as representing degradation level of ongoing calls at a random point of time in the perspective of systems's view. From the service users perspective, however, the degradation level should be defined considering the period of total service time because the service users evaluate a system's service on the basis of their degradation experiences during call durations. DPR which represents the average portion of a call's life time when it is degraded is a meaningful measure from the service user's standpoints. However, in case of multiple bandwidth levels where the number adaptable bandwidth is more than three, the degradation depth of ongoing calls cannot be represented only with the information of periods such as DPR.

Therefore, in this paper, new QoS parameters, Degradation Degree Ratio (DDR) and Degradation Area Ratio (DAR) are proposed to represent the service user's degradation in multi-level bandwidth adaptation situation.

An analytical model for estimating DPR proposed by Kwon *et al*. (2003) assumes that the degradation probability of an observed call is constant and independent of system states. As a result, it does not properly represent a real situation in that the degradation probability of an observed call is variable dependent on the system states and the types of BAAs. Therefore, in order to represent the real situation practically, we compose state variables as the number of calls in the system and the bandwidth layer of an observed call.

Our contributions are summarized as follows:

First, new QoS parameters, Degradation Degree Ratio (DDR) and Degradation Area Ratio (DR) are defined from the service user's perspective in multi-level bandwidth adaptation systems. DDR and DAR can be meaningful QoS measures together with an existing QoS parameter, DPR, if the revenue or cost rate is determined in proportion to the assigned bandwidth size to a call.

Second, a two-dimensional continuous time Markov chain (CTMC) with an absorbing state is composed to capture the variation in the length of a call's lifetime and its state transitions practically. Using this model, the QoS measures such as DPR, DDR and DARcan be calculated.

The proposed model can be used to design optimal threshold-type CAC parameters and analyze the performance of some BAAs.

2. Model Assumption and System Description

2.1 Model assumption

An isolated cell approach is adopted on the assumption that the arrival and departure patterns of traffic are statistically identical in all neighboring cells and the new call arrival rate and the handoff arrival rate into a cell concerned are measured statistically. The total bandwidth capacity in each cell is the same and is denoted by C assuming a fixed channel allocation scheme. Discrete bandwidth values are considered assuming layered coding technique. A set of bandwidth is expressed as $B=\{b_1,b_2,\cdots,b_I\}$ where $b_i< b_{i+1}$ for $i=1\cdots, I-1$. A layer 1 indicates the lowest service level and I indicates the highest service level. $N_{\max}$ denotes the maximum number of calls a system can support by allocating only a minimum bandwidth b_1 and can be calculated as $\lfloor C/b_1 \rfloor$.

As is usual in the literature, it is assumed that call arrivals to a cell form a Poisson process. Let λ^{NC} and λ^{HC} be the new call arrival rate and the handoff call arrival rate respectively. The call holding time is assumed to follow an exponential distribution with a mean, $1/\mu$. The amount of time that a user stays in a cell before handoff is also assumed to follow an exponential distribution with a mean, $1/\eta$. Then the cell departure rate of a call is $\nu=\mu+\eta$. Note that a departure event is regarded as the event a user terminates a call or crosses over to another cell region (handoff). The traffic modeling assumption is depicted in <Figure 1>.

In order to estimate the QoS measure strictly from the service user's perspective, one incoming call is chosen at random time and its state transitions are observed until it departs from a cell. Hereafter this is called a tagged call. Let the state of the system be defined as $S=(i,\ n)$ for $i=\{1.\cdots,I\}$ and $n\in\{1,\cdots,N_{\max}\}$. Here i indicates the bandwidth layer in which a tagged call is

included and n indicates the total number of ongoing calls in a cell including a tagged call. The value of system state vector S then changes according to CTMC with one absorbing state. The system enters the absorbing state when a tagged call departs from the cell.

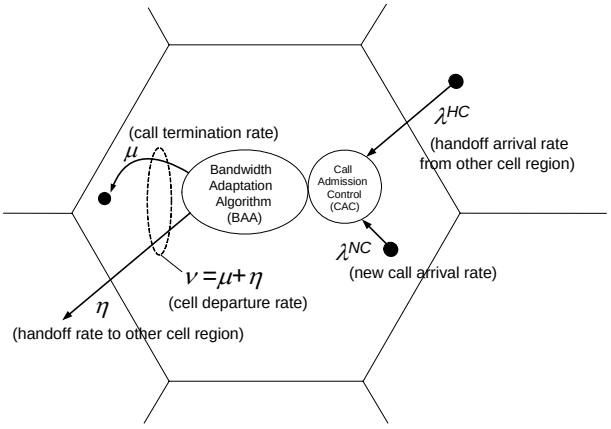

Figure 1. Traffic modeling assumption

2.2 Call Admission Control

To give priority to handoff calls, this study adopted a threshold-type CAC algorithm where a newly arriving call is blocked if the current number of ongoing call is equal to or greater than threshold value N_{th}. On the other hand, an incoming handoff call is accepted regardless of the number of ongoing calls if the call can be accommodated. This threshold-type CAC has been demonstrated to be best in terms of efficiency and applicability under single-class situation (Guerin, 1988; Ramjee *et al*. 1996).

An arrival rate function when there are n calls in a cell, $\lambda(n)$, can be given as follows.

$$\lambda(n) = \begin{cases} \lambda^{HC} + \lambda^{NC} & , \quad 0 \leqq n \leqq N_{th} \\ \lambda^{HC} & , \quad N_{th} \leqq n < N_{\max} \\ 0 & , \quad o/w \end{cases} \tag{1}$$

The call dwelling time in a cell is assumed to follow an exponential distribution with mean $1/\nu$, where ν is the departure rate in a given cell. A departure rate function when there are n calls in a cell is $\nu(n) = (n-1) \cdot \nu$ for $n \in \{1, \cdots, N_{\max}\}$. The term $(n-1)$ means the number of users in a cell excluding a tagged call.

2.3 The Bandwidth Adaptation Algorithms

BAA will be triggered whenever there are arrival acceptance events and call departure events. Various adaptation algorithms are applicable to the model, however, the following two typical adaptation algorithms are adopted in this study.

2.3.1 Bandwidth Adaptation Algorithm for Fairness (BAA-F)

Fairness for the ongoing calls is always ensured in BAA-F algorithm either by upgrading the calls from the lowest bandwidth layer when there is an available capacity or degrading calls from the highest bandwidth layer when there is a need for bandwidth. The calls are selected one by one at each step. The adaptation algorithm allows calls to move only to an adjacent layer at each adaptation event in order to prevent users from experiencing severe fluctuations. Once a call moves to an adjacent layer, it can only move again after all other calls from this layer have moved to a new layer.

2.3.2 Bandwidth Adaptation Algorithm for Minimizing the Number of Degraded calls (BAA-MND)

BAA-MND algorithm aims to minimize the number of calls whose bandwidths are lower than target bandwidth, b_{tar}, at any time. BAA-MND algorithm should allocate a minimum bandwidth to an incoming call from the possible bandwidth set and reallocate the bandwidth of the remaining calls according to the predetermined numbers of calls at each bandwidth layer. In the case of arriving calls, the algorithm degrades the calls according to the order of the size of the bandwidth. It attempts to upgrade the calls whose bandwidths are lower than b_{tar} to the level of b_{tar} or more when there is an outgoing handoff call or a call completion. Calls with the same bandwidth are selected in a random order.

3. QoS Parameters in Adaptive System

This study follow the definition of DPR proposed by Kwon *et al*. (2003) and introduce new definitions of DDR and DAR in the proposed analytic model.

DPR represents the average portion of a call's lifetime where a call is allocated a bandwidth, which is lower than the target bandwidth, b_{tar}. Let τ be a call's lifetime in a cell and τ_i be the total length of time that the call has been allocated bandwidth of b_i during τ. If $\tau_d \left(= \sum_{i=1}^{tar-1} \tau_i \right)$ is the total degradation period, then

DPR = τ_d/τ.

DDR represents the average ratio of degree of degradation during a total degradation period, τ_d. If the degradation degree of each bandwidth layer is $\triangle b_i = Max\{0, b_{tar} - b_i\}$, then DDR $= \sum_{i=1}^{tar-1} \frac{\triangle b_i}{\triangle b_1} \frac{\tau_i}{\tau_d}$.

DAR represents the average ratio of a call's degradation level considering both the period and the degree of degradation and is calculated as a product of DPR and DDR. So, DAR $= \sum_{i=1}^{tar-1} \frac{\triangle b_i}{\triangle b_1} \frac{\tau_i}{\tau}$.

For example, if DPR = 0.2 and DDR = 0.3, then this indicates that a call generally experiences an average of 30% bandwidth degradation to the worst degradation case during the 20% period of its lifetime. In this case DAR = 0.06, which implies a call undergoes an average of 6% bandwidth degradation during its lifetime.

4. Estimation of DPR, DDR and DAR

Procedures for calculating the QoS parameters, DPR, DDR and DAR are presented in this section. First, the number of calls at each layer is determined when the total number of calls in a system is given. The state transition probabilities and a state transition matrix are calculated. The tagged call's expected cumulative times spent at each bandwidth layer is then derived using this transition matrix and Kolmogorov's equations. Finally, DPR, DDR and DAR formulas are presented.

4.1 Calculation of the Number of Calls at Each Layer

Let the vector $L(n) = (l_1(n), \cdots, l_i(n), \cdots, l_I(n))$ represent the distribution of calls at each layer and $l_i(n)$ be the number of calls at layer i when there are a total of n calls in a given cell. Note that the vector $L(n)$ can be uniquely calculated according to BAA.

4.1.1 $L(n)$ of BAA-F

BAA-F algorithm ensures fairness for ongoing calls and only one or two layers can have calls among all the bandwidth layers. If there are two layers with calls, they must be adjacent to each other. When $n > \lfloor C/b_I \rfloor$, the layers with calls can be denoted by i^* or $i^* - 1$ for $i^* \in \{2, \cdots, I\}$ and can be determined by finding maximum integer value i satisfying equation (2) for $\exists x \ s.t. \ x \in \{0, \cdots, n-1\}$.

$$C = b_{i-1} \cdot x + b_i \cdot (n - x) + \delta \quad (0 \leq \delta < b_{i-1}), \tag{2}$$

After the layers with calls (i^*) are determined, the numbers of calls of those layers can be determined by searching the minimum integer value x, s.t. $x \in \{0, \cdots, n-1\}$ satisfying equation (2). For example, if $C = 20$, $B = (2,5,7,10)$, then $L(n)$ for $n \in \{1, \cdots, 10\}$ can be obtained by

$$\begin{aligned}\{L(n) | n = 1, \cdots, 10\} &= \{ (0,0,0,1), (0,0,0,2), (0,1,2,0), (0,4,0,0), \\ &\quad (2,3,0,0), (4,2,0,0), (5,2,0,0), \\ &\quad (7,1,0,0), (9,0,0,0), (10,0,0,0)\}\end{aligned}$$

4.1.2 $L(n)$ of BAA-MND

BAA-MND algorithm seeks to minimize the number of calls whose bandwidths are lower than target bandwidth. If $n \cdot b_{tar} < C$ then as much bandwidth as possible is assigned to each call from the maximum bandwidth b_I to the target bandwidth b_{tar} in descending order.

If $n \cdot b_{tar} \geq C$ then the minimum non-negative integer value x satisfying equation (3) is determined.

$$C = (n - x) \cdot b_{tar} + x \cdot b_1 + \delta, \quad (0 \leq \delta < b_{tar} - b_1) \tag{3}$$

It can be seen that x^* is the minimum number of degraded calls. b_{tar} is assigned to $n - x^*$ calls and as much bandwidth as possible is assigned to each call from bandwidth b_{tar-1} to the minimum bandwidth b_1 in descending order with the remaining capacity, $C - (n - x^*) \cdot b_{tar}$.

For example, if $C = 20$, $B = (2,5,7,10)$ and $b_{tar} = 7$, then $L(n)$ for $n \in \{1, \cdots, 10\}$ can be obtained by

$$\begin{aligned}\{L(n) | n = 1, \cdots 10\} &= \{ (0,0,0,1), (0,0,0,2), (0,1,2,0), \\ &\quad (2,0,2,0), (3,0,2,0), (4,1,1,0), \\ &\quad (6,0,1,0), (7,1,0,0), (9,0,0,9), (10,0,0,0)\}\end{aligned}$$

4.2 States Transition Probabilities

The state transition probabilities are calculated in this section.

<Figure 2> shows the output rate of CTMC when the state of a tagged call is given as $S = (i, n)$

Let $P_{(i,n)(i',n')}$ denote the probability that a tagged call moves from state (i, n) to state (i', n') when an arrival or departure event occurs. For simplicity, a new

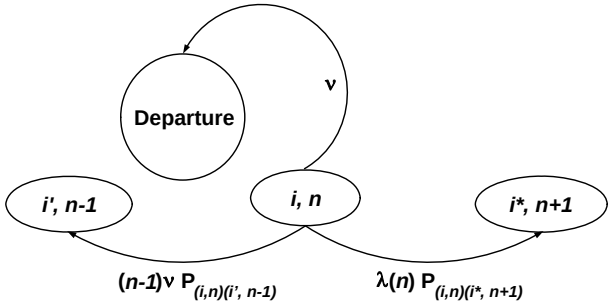

Figure 2. Output rate of a tagged call from $S=(i,n)$

vector $L^k(n)$, which represents the distribution of calls at each bandwidth layer after one call at layer *k*, is excluded from the vector $L(n)$, is introduced. $L^k(n)=(l_1^k(n),\cdots,l_I^k(n))$ can be expressed simply as $(l_1(n),\cdots,l_k(n)-1,\cdots,l_I(n))$. In the case of a call's departure, n' becomes $n-1$. The probability $P_{(i,n)(i',n-1)}$ can be calculated as follows.

$$P_{(i,n)(i',n-1)}=\sum_{k=1}^{I}P_k^{(i,n)(i',n-1)}\cdot P_k^{Dep}(i,n) \qquad (4)$$

In equation (4), $P_k^{Dep}(i,n)$ is the probability that a call except a tagged one departs from layer *k* when the system state is (*i* ,*n*). This is calculated using equation (5).

$$P_k^{Dep}(i,n)=\begin{cases}\dfrac{l_k(n)-1}{n-1} & n\geq 2,\ k=i\\ \dfrac{l_k(n)}{n-1} & n\geq 2,\ k\neq i\\ 0 & n=1\end{cases} \qquad (5)$$

$P_k^{(i,n)(i',n-1)}$ is the conditional probability that a tagged call is shifted from layer i to la*yer* i' conditioned that a departure event occurs at bandwidth layer *k*. It can be calculated by equation (6).

$$P_k^{(i,n)(i',n-1)}=\begin{cases}\left(\dfrac{Max\{0,l_i^k(n)-l_i(n-1)\}}{l_i^k(n)}\right)\cdot\\ \left(\dfrac{Max\{0,l_{i'}(n-1)-l_{i'}^k(n)\}}{\sum_{s=1}^{i}Max\{0,l_s(n-1)-l_s^k(n)\}}\right), & i\neq i'\\ 1-\dfrac{Max\{0,l_i^k(n)-l_i(n-1)\}}{l_i^k(n)}, & i=i'\end{cases} \qquad (6)$$

In equation (6), when $i\neq i'$, the first term of RHS is the probability that a tagged call is randomly selected to be adapted among $l_i^k(n)$ calls at layer i. The second term of RHS denotes the probability that a call for adaptation at layer i moves to the layer i' among all layers. When $i=i'$, the event that a tagged call remains at layer i is complementary of the event that a tagged call is selected to be adapted at layer i.

In the case of a call's arrival, n' becomes $n+1$. When a call has arrived, it should be allocated the smallest bandwidth at the layer k in order to minimize the degree of bandwidth degradation of the ongoing calls. The probability $P_{(i,n)(i',n+1)}$ can then be calculated as follows.

$$P_k^{(i,n)(i',n+1)}=\begin{cases}\left(\dfrac{Max\{0,l_i^k(n)-l_i(n+1)\}}{l_i^k(n)}\right)\cdot\\ \left(\dfrac{Max\{0,l_{i'}(n+1)-l_{i'}^k(n)\}}{\sum_{s=1}^{i}Max\{0,l_s(n+1)-l_s^k(n)\}}\right), & i\neq i'\\ 1-\dfrac{Max\{0,l_i^k(n)-l_i(n+1)\}}{l_i^k(n)}, & i=i'\end{cases} \qquad (7)$$

In equation (7), when $i\neq i'$, the first term of RHS is the probability that a tagged call is randomly selected to be adapted at layer i. The second term of RHS denotes the probability that a call for adaptation at layer i moves to layer i' among all layers. When $i=i'$, the event that a tagged call remains at layer i is complementary of the event that a tagged call is selected to be adapted.

4.3 Generation of Transition Matrix

Let Q denote the infinitesimal generator matrix of CTMC in our system. It can be expressed as follows.

$$Q=\begin{bmatrix}D & \hat{Q}\\ 0 & 0\end{bmatrix} \qquad (8)$$

The component D is a column vector of which elements are all ν. It represents the transition rate in which a tagged call enters an absorbing state from all other states. $\hat{Q}$ denotes the sub matrix of Q pertaining to the states that are not absorbing states. $\hat{Q}^{ii'}$ denotes the sub matrix of $\hat{Q}$ which represents a transition of a tagged call from the bandwidth layer i *to* i'. The elements of $\hat{Q}^{ii'}$ are obtained as follows.

$$\hat{q}^{ii'}(n,n')=\begin{cases}\nu(n)\cdot P_{i,n)(i',n')}, & n'=n-1\\ -(\nu(n)+\lambda(n)), & i=i',n=n'\\ \lambda(n)\cdot P_{(i,n)(i',n')}, & n'=n+1\\ 0, & o/w\end{cases}$$

$$for\ i,i'=1,\cdots I \text{ and } n,n'=1,\cdots N_{\max} \qquad (9)$$

Let the vector of state probabilities at time t be denoted as $\pi(t)$ and the sub vector of $\pi(t)$ excluding an absorbing state be denoted by $\hat{\pi}(t) = [\hat{\pi}_S(t)]$.

Let the vector $Y = \{y_s\}$ and $y_S = \lim_{t\to\infty} \int_0^t \pi_S(\tau) d\tau$. A vector-matrix form can be obtained as $Y \cdot \hat{Q} = -\pi(0)$. Here $\pi(0)$ is the vector of the initial state probability. y_s is the expected time the CTMC spends in a state S until a tagged call terminates (Trivedi, 2002).

4.4 Calculation of Layer Residence Ratio i (LRR^i)

Layer Residence Ratio $i(LRR^i)$ is defined as the average dwelling time of a call at layer i during its lifetime in a cell. By calculating LRR^i for $i = 1, \cdots I$, DPR and DDR can be obtained for a target bandwidth layer. First, a conditional $LRR^i_{(i_0,n_0)}$ should be obtained when an initial state of a tagged call is given as $S = (i_0, n_0)$. The conditional $LRR^i_{(i_0,n_0)}$ is given by equation (10).

$$LRR^i_{(i_0,n_0)} = \frac{\sum_{n=1}^{N_{\max}} y_{(i,n)|(i_0,n_0)}}{\frac{1}{\nu}} = \nu \cdot \sum_{n=1}^{N_{\max}} y_{(i,n)|(i_0,n_0)} \quad (10)$$

As mentioned above, an observing call is chosen at random time. In order to assign an initial state of the call, the number of calls in a system at random time needs to be known.

A one-dimensional state variable n is defined as the number of calls in a system in order to derive a tagged call's initial state probabilities. A birth-death process can then be composed with the state space $\{0. \cdots,. N_{\max}\}$ The state transition diagram is given in <Figure 3>.

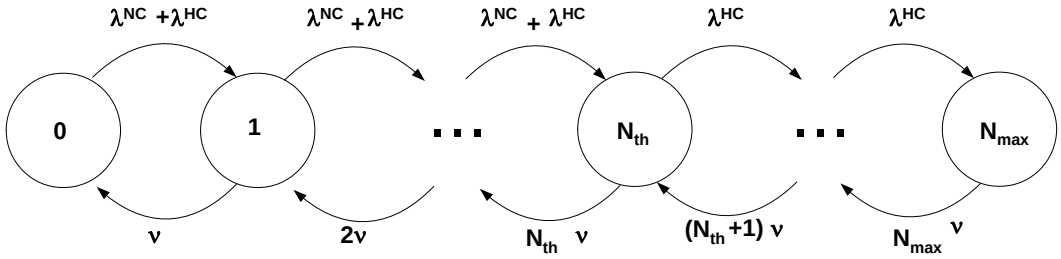

Figure 3. State transition diagram of cell concerned

The limiting probabilities of this CTMC, π_i, can be determined as follows (Kleinrock, 1975).

$$p_i = \begin{cases} \frac{(\lambda^{NC} + \lambda^{HC})^i}{i!\nu^i} p_0 & , \ 1 \le i < N_{th} \\ \frac{(\lambda^{NC} + \lambda^{HC})^{N_{th}} (\lambda^{HC})^{i - N_{th}}}{i!\nu^i} p_0 & , \ N_{th} \le i \le N_{\max} \end{cases} \quad (11)$$

$$p_0 = \left[1 + \sum_{i=1}^{N_{th}-1} \frac{(\lambda^{NC} + \lambda^{HC})^i}{i!\nu^i} + \sum_{i=N_{th}}^{N_{\max}} \frac{(\lambda^{NC} + \lambda^{HC})^{N_{th}} (\lambda^{HC})^{i - N_{th}}}{i!\nu^i} \right]^{-1} \quad (12)$$

The probability of a tagged call's initial state, $p_{(i,n)}$, can be defined using equation (13) and equation (14).

$$p_{(i,n)} = \begin{cases} p_{n-1} & i = k \\ 0 & o/w \end{cases} \quad for \ i \in \{1, \cdots I\}, \ n \in \{1, \cdots, N_{\max}\} \quad (13)$$

In equation (13), *k* represents the bandwidth layer having a call (calls) with the smallest bandwidth from the vector $L(n)$. LRR^i is finally obtained in the form of an expected average using the initial state probabilities as weights.

$$LRR^i = \sum_{n=1}^{N_{\max}} LRR^i_{(i,n)} \cdot p_{(i,n)} \quad (14)$$

4.5 Calculation of DPR, DDR and DAR

Let tar be the target bandwidth layer, then DPR, DDR and DAR can finally be calculated as follows.

$$DPR = \sum_{i=1}^{tar-1} LRR^i \quad (15)$$

$$DDR = \sum_{i=1}^{tar-1} \left(\frac{b_{tar} - b_i}{b_{tar} - b_1} \cdot \frac{LRR^i}{DPR} \right) \quad (16)$$

$$DAR = DPR \cdot DDR \quad (17)$$

5. Numerical Results

The characteristics of the estimated DPR, DDR and DAR as the QoS parameters in an adaptive system are demonstrated in this section. Comparisons of performances of BAA-F and BAA-MND are also illustrated.

The total bandwidth capacity of a cell is $C = 100$(in basic bandwidth units). The bandwidth set, $B = (2,5,7,10)$ and the number of layers, $I = 4$. Here it can be seen that $N_{\max} = 50$. A target bandwidth, $b_{tar} = 7$. The call duration time is $1/\mu = 500$(sec). The handoff arrival rate ratio to the total call arrival rate is assumed to be 1:2, i.e., $\lambda^{HC}/(\lambda^{HC}+\lambda^{NC})$ = 1/3 as is usual in the literature (Lin *et al*. 1994).

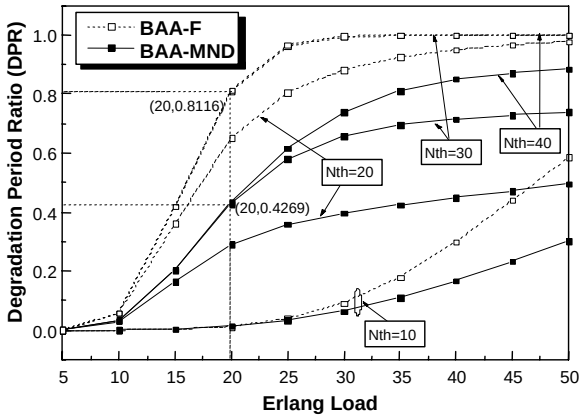

Figure 4. Degradation Period Ratio (DPR) with different thresholds

<Figure 4> shows DPR of BAA-F and BAA-MND as the traffic load, which is measured in Erlangs, increases. It is shown that DPR of BAA-F is notably higher than that of BAA-MND at the same threshold valuein the entire traffic load. For example, when the Erlang load = 20 and the threshold value of CAC = 30, it is observed that DPR of BAA-F is 0.8116 and that of BAA-MND is 0.4269. These QoS values indicate that a user generally experiences bandwidth degradation during an 81% period of the call's lifetime if BAA-F is applied and a 43% period of the call's lifetime if BAA-MND is applied under such system conditions.

<Figure 5> shows DDR of both BAA algorithms as the traffic load increases. It can be seen that DDR of BAA-F is notably lower than that of BAA-MND at the same threshold value in the entire traffic load. For example, when the Erlang load = 20 and the threshold value of CAC = 30, it is observed that DDR of BAA-F

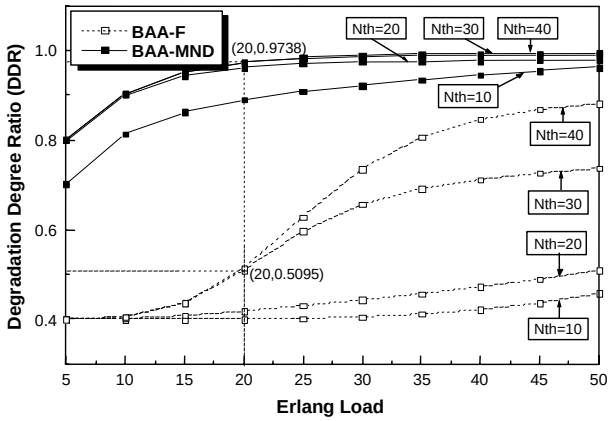

Figure 5. Degradation degree ratio (DDR) with different thresholds

is 0.5095 and that of BAA-MND is 0.9738. These QoS values suggest that a user generally experiences an average of 51% bandwidth degradation to the worst degradation case during the degradation period of the call's lifetime if BAA-F is applied and 97% if BAA-MND is applied under such system conditions.

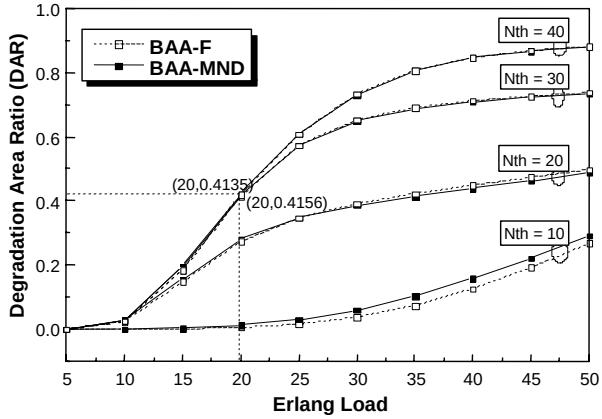

Figure 6. Degradation area ratio (DAR) with different thresholds

<Figure 6> shows DAR of both BAA algorithms as the traffic load increases. The results show that DAR of BAA-F andBAA-MND are almost the same particularly in high threshold values. For example, when the Erlang load = 20 and the threshold value of CAC = 30, DAR of BAA-F is 0.4135 and that of BAA-MND is 0.4157. These QoS values indicate that a user generally undergoes an approximately 41% bandwidth degradation during a call's lifetime regardless of the applied BAA algorithms under such system conditions.

This phenomenon is due to the difference in the objectives of the two BAA algorithms. BAA-F algorithm attempts to share the total bandwidth degradation to all

ongoing calls but BAA-MND algorithm tries to concentrate the total bandwidth degradation on the minimum number of calls, so the degradation probability and the period of each call become larger and the degradation degree becomes smaller when BAA-F algorithm is applied.

From the results of Figure 4,5 and 6, it is concluded that the types of applied BAA do not significantly affect the total bandwidth degradation of a call. However, the types of BAA mainly influence the bandwidth degradation period and degree.

<Figure 7> shows New call Blocking Probability (NBP) and HDP versus the Erlang load in the same experiment. As illustrated, HDP is almost zero (lower than 9.5E-6) in the entire traffic load and NBP increases when the traffic load increases.

<Figure 8> shows DPR of both BAA algorithms as the threshold value of CAC increases. It reveals that DPR increases as the threshold value increases because more new calls are accepted into the system. It can be observed that DPR of BAA-F is notably higher than that of BAA-MND at the same Erlang load in the entire threshold value.

<Figure 9> shows DDR of both BAA algorithms as the threshold value increases. It reveals that DDR increases as the threshold value increases because more new calls are accepted into the system. It can be observed that DDR of BAA-F is notably lower than that of BAA-MND at the same Erlang load in the entire threshold value.

<Figure 10> shows DAR of both BAA algorithms as the threshold value of CAC increases. It can be seen that DAR of BAA-F and BAA-MND are almost the same particularly in the high threshold value of CAC.

<Figure 11> shows that HDP is almost zero (lower than 3.12E-4) in the entire threshold value. It can be observed that NBP decreases as the threshold value increases because more new calls can be accepted into the system.

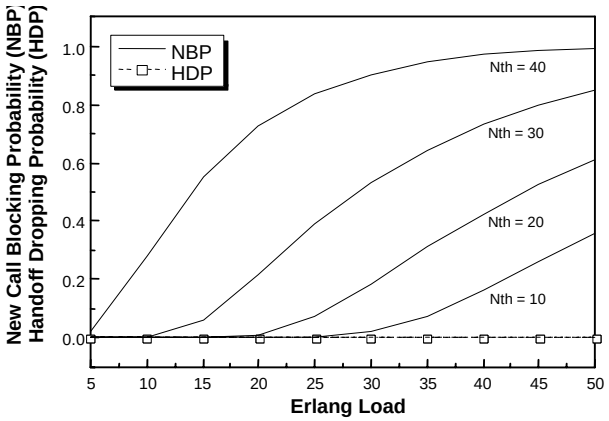

Figure 7. Blocking & dropping probability with different thresholds

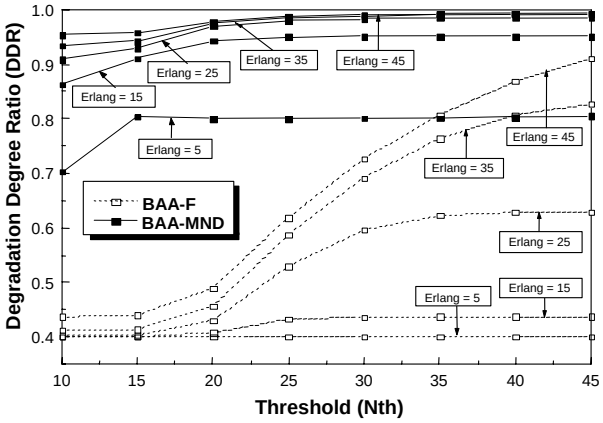

Figure 9. Degradation degree ratio (DDR) with different Erlang load

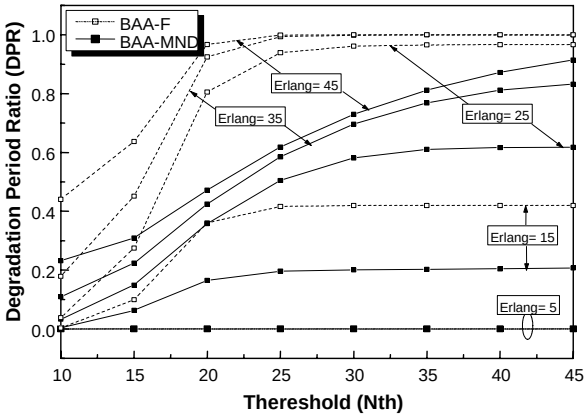

Figure 8. Degradation period ratio (DPR) with different Erlang load

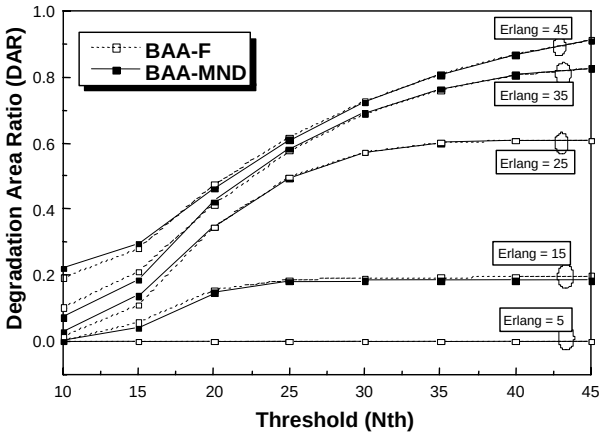

Figure 10. Degradation area ratio (DAR) with different Erlang load

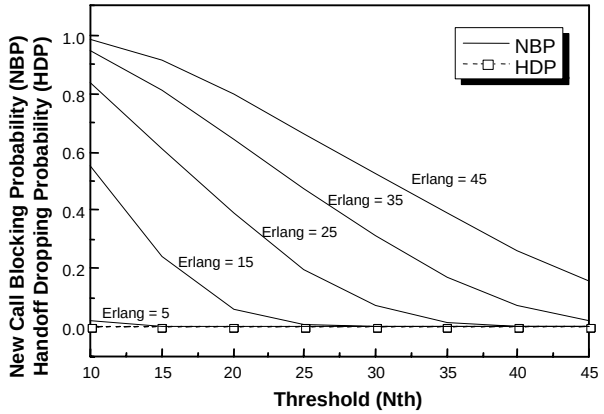

Figure 11. Blocking and dropping probability with different Erlang load

In <Figure 12> and <Figure 13>, the optimal threshold values of BAA-F and BBA-MND are the same at 19 when the Erlang load is 15 and NBP is set to be lower than 0.1. Under these system conditions, DAR of BAA-F and BAA-MND are almost similar (0.1301 and 0.1387, respectively). DPR of BAA-F and BAA-MND algorithms are 0.3197 and 0.1487, and DDR are 0.4072 and 0.9355 respectively. This means that the system utilization or the total bandwidth degradation of both adaptation algorithms is not different. However, BAA-F algorithm provides a longer degradation period and a smaller degradation degree than BAA-MND algorithm.

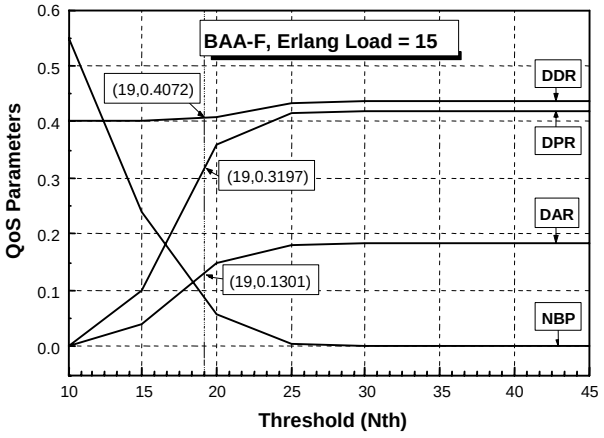

Figure 12. QoS parameters of BAA-F under optimal threshold

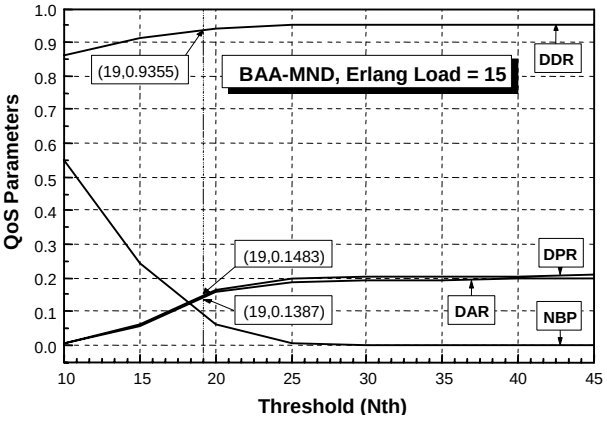

Figure 13. QoS parameters of BAA-MND under optimal threshold

6. Conclusions

We propose DDR and DAR as new QoS parameters. DDR and DAR can be meaningful QoS measures if the revenue or cost rate is determined in proportional to the assigned bandwidth size to a call. An analytic model is also proposed to estimate DPR, DDR and DAR from the service user's perspective in adaptive mobile cellular networks. A threshold-type CAC and two BAAs, which guarantees fairness (BAA-F) and minimizes the number of degraded calls (BAA-MND) are adopted in this study. Numerical examples show that the types of adopted bandwidth adaptation algorithms produce similar system utilization and give the same total bandwidth degradation. The proposed model is applicable to the design of an optimal threshold-type CAC policy and the performance analysis of each adopted BAA.

References

Ahn, K. M. and Kim, S. (2003), Optimal bandwidth allocation for bandwidth adaptation in wireless multimedia networks, *Computers & Operations Research*, **30**(13), 1917-1929.

Bharghavan, V., Lee, K. W., Lu, S. W., Ha, S. W., Li, J. R. and Dwyer, D. (1998), The timely adaptive resource management architecture. *IEEE Personal Communications Magazine*, **5**(4), 20-31.

Chou, C. T. and Shin, K. G. (2004), Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service, *IEEE Transactions on Mobile Computing*, **3**(1), 5-17 .

Guerin, R.(1988), Queueing-blocking system with two arrival streams and guard channels, *IEEE Transactions on Communications*, **36**(2), 153-163.

Jain, R. and Knightly, E. W. (1999), A framework for design and evaluation of admission control algorithms in multi-service mobile networks, *Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, **3**, 1027-1035.

Kwon, T., Choi, Y., Bisdikian, C. and Naghshineh, M. (1998), Call admission control for adaptive multimedia in wireless/ mobile networks, *Proceedings of ACM Workshop on Wireless Mobile Multimedia.* 111-116.

Kwon, T., Choi, Y., Bisdikian, C. and Naghshineh, M. (1999), Measurement-based call admission control for adaptive multimedia in wireless/mobile networks, *Proceedings of IEEE Wireless Communications and Networking Conference.* 540-544.

Kwon, T., Park, I., Choi, Y. and Das, S. (1999), Bandwidth adaptation algorithms with multi-objectives for adaptive multimedia services in wireless/mobile networks, *Proceeding of ACM Workshop on Wireless/Mobile Multimedia,* 51-58.

Kwon, T., Choi, Y., Bisdikian, C. and Naghshineh, M. (2003), QoS provisioning in wireless/mobile multimedia networks using an adaptive framework, *Wireless Networks*, **9**(1), 51-59.

Naghshineh, M. and Willebeek-LeMair, M. (1997). End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework, *IEEE Communications Magazine*, **35**(11), 72-81.

Nasser, N. and Hassanein, H. (2004), Connection-level performance analysis for adaptive bandwidth allocation in multimedia wireless cellular networks, *2004 IEEE International Conference on Performance, Computing, and Communications*, 61-68.

Ramjee, R., Nagarajan, R. and Towsley, D. (1996), On optimal call admission control in cellular networks *INFOCOM'96. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, **1**, 43-50.

Ross, S. M. (2000), *Introduction to Probability Models*, 7th ed., San Diego : Academic Press.

Trivedi, K. S. (2002), *Probability and Statistics with Reliability, Queuing and Computer Science Applications,* 2nd ed., New York : John Wiley and Sons.

Wang, J., Li, M., Yang, X. and Huang, Z. (2001), Utility-based call admission control for adaptive mobile services, *Proceedings of International Conference on Computer Networks and Mobile Computing*, 91-96.

Xiao, Y., Chen, C. L. P. and Wang, Y. (2000), Quality of service and call admission control for adaptive multimedia services in wireless/mobile networks, *Proceedings of the IEEE National Aerospace and Electronics Conference,* 214-220.

Xiao, Y., Chen, C. L. P. and Wang, B. (2002), Bandwidth degradation QoS provisioning for adaptive multimedia in wireless/mobile networks, *Computer Communications,* **25**(13), 1153-1161.

Zaruba, G. V., Chlamtac, I. and Das, S. K. (2002), A prioritized real-time wireless call degradation framework for optimal call mix selection. *Mobile Networks and Applications,* **7**, 143-151.

Zhao, P. and Zhang, H. M. (2001), A new CAC algorithm for adaptive service in mobile network, *Proceedings of International Conferences on Info-tech and Info-net*., **2**, 199-204. Beijing.