A New Joint Packet Scheduling/Admission Control Framework for Multi-Service Wireless Networks

Fei Long, Gang Feng, and Junhua Tang

Abstract: **Quality of service (QoS) provision is an important and indispensable function for multi-service wireless networks. In this paper, we present a new scheduling/admission control framework, including an efficient rate-guaranteed opportunistic scheduling (ROS) scheme and a coordinated admission control (ROS-CAC) policy to support statistic QoS guarantee in multi-service wireless networks. Based on our proposed mathematical model, we derive the probability distribution function (PDF) of queue length under ROS and deduce the packet loss rate (PLR) for individual flows. The new admission control policy makes admission decision for a new incoming flow to ensure that the PLR requirements of all flows (including the new flow) are satisfied. The numerical results based on ns-2 simulations demonstrate the effectiveness of the new joint packet scheduling/admission control framework.**

Index Terms: **Admission control, packet scheduling, quality of service (QoS), wireless networks.**

I. INTRODUCTION

The traffic control mechanisms designed for wireline networks, such as packet scheduling scheme and admission control policy, are not directly applicable to wireless networks since their design has not taken into account some unique problems in wireless networks, such as varying wireless link capacities, location-dependent and bursty bit errors, joint scheduling of uplink and downlink flows, power constraint in wireless cells, etc. Specifically, three characteristics in wireless signal propagation determine the quality of the wireless channel: Path loss variation, slow log-normal shadowing, and fast multipath fading[1]. As a result, packet scheduling in wireless networks should also be link state dependent. All these make it a challenging issue to design a good traffic control scheme in wireless networks.

Quality of service (QoS) provisioning for on-line traffic such as a conversational call or a video conferencing call is an indispensable function of multi-service wireless networks. These on-line traffic flows have multiple QoS requirements, such as packet delay bound and packet loss rate (PLR), etc. These QoS requirements are too multifold to be satisfied by a single scheme. A traffic control framework, including various mechanisms such as packet scheduling and admission control, is necessary for meeting multiple QoS requirements. In this paper, we focus on a joint packet scheduling/admission control framework for providing statistic QoS guarantees in multi-service wireless networks. We first review a novel packet scheduling scheme we proposed for wireless networks, called rate-guaranteed opportunistic scheduling (ROS)[2]. ROS can guarantee statistical delay bound for flows with arrival curve constraint. Furthermore, ROS can achieve optimal throughput and link utilization while providing fairness guarantee between flows. Both channel states and fairness are considered in the computation of priorities of flows, based on which flows are scheduled. The fairness degree of each flow is represented by the differences between "relative throughput" of each flow and the maximum "relative throughput" value among all flows. The proof of optimality of ROS and the study on its behaviors are reported in a separate paper[2]. In this paper, we focus on a new admission control mechanism which works together with ROS to provide statistical QoS guarantees for multi-service wireless networks.

Admission control is another essential building block of the traffic control framework for QoS provisioning besides packet scheduling. The main function of admission control scheme is to decide the admission of a new incoming flow by judging whether the QoS of existing flows is violated and QoS of the new incoming flow can be satisfied or not. Recently, a lot of admission control schemes have been proposed for wireless networks[3]–[5]. These schemes deduce different admission regions as they use different admission criteria. Typical admission criteria for wireless networks include the signal interference ratio (SIR)[3], throughput[4], packet loss rate[5], etc.

Packet loss rate and delays are two important QoS parameters for real-time communication services. However, voice and/or video users can usually tolerate the loss of a very small fraction of packets. Thus, it is possible to make a more efficient utilization of network resources by providing statistical QoS guarantees. In other words, the traffic control framework operates the network in such a way that congestion phenomena, which leads to packet losses or delays, only occurs with predefined small probabilities. The packet losses in modern wireless networks mainly consist of two parts: Packet losses during transmission and packet losses due to buffer overflow in case of congestion at the medium access control (MAC) layer. In modern high-speed wireless networks, the packet losses due to erroneous transmission are well controlled by some mechanisms such as error control schemes and loss recovery and channel states prediction, and occur with a smaller probability than that due to congestion[5].

Traditionally, admission control policy for wireless networks is designed based on the criteria of SNR[3] or throughput[4]. A joint scheduling/admission control framework is proposed in[3] to optimize resource management for uplink in wideband code-division multiple-access (WCDMA). New variable bit-rate (VBR) user is admitted only if there is spare capacity after the minimum carrier to total interference ratio (C/I) requirement of the new user is satisfied. A traffic aided smooth admission control (SAC) for opportunistic communication systems, which can provide throughput guarantee, has been pro-

Manuscript received July 25, 2005; approved for publication by Urbashi Mitra, Guest Editor, November 11, 2005.

The authors are with the School of EEE, Nanyang Technological University, Singapore, email: longf@pmail.ntu.edu.sg, {egfeng, ejhtang}@ntu.edu.sg.

posed in[4]. SAC spreads the decision over a trial period, during which the amount of time resources allocated to new incoming users increase gradually. The throughput in this trial period is measured and used to make admission decision.

There are also some proposed admission control policies using packet loss rate (PLR) as design criteria. An edge-to-edge measurement based admission control[5] is proposed to maintain PLRs for voice flows of both directions. Lack of mathematical modeling is its shortcoming. In a new data traffic control scheme proposed for integrated service code-division multiple-access (CDMA) systems[6], joint admission/congestion control is used to maintain the packet error rate (PER) of real-time traffic while non-real-time traffic can use the residual link resources. The admission control policy only admits the data call which will not experience higher delay than the required QoS constraint. The congestion control scheme ensures that the PER requirement of real-time voice traffic is satisfied. Markovian model is used to evaluate PER, blocking probability, throughput, and delay.

To provide statistical QoS guarantees, we propose a new efficient admission control policy, which uses PLR as the admission criteria to work in conjuction with ROS in multi-service wireless networks. PLR is an important QoS parameter for real-time traffic. Using PLR as a criteria for our admission control policy design provides us an effective and efficient approach for QoS provisioning. By investigating the behaviors of packet queues in ROS, we deduce the mathematical expression of probability distribution of queue length. After simplification and approximation, we can obtain probability distribution function (PDF) results similar to $M/D/c$ model. Based on the properties of PDF, we design a new admission control policy called ROS-CAC specially for ROS scheduler. The simulation results based on ns-2[7] demonstrate the effectiveness of our new admission control policy.

The remainder of this paper is organized as follows. In Section II, we introduce the system model. In Section III, we provide a review for ROS. After that, we analyze the PLR performance of ROS and present our new admission control policy in Section IV. We examine the behaviors and performance of our proposed admission control policy via simulation in Section V and conclude the paper in Section VI.

II. SYSTEM MODEL

We consider a time-slotted system, in which time-division multiplexing (TDM) is deployed, consisting of one base station and many mobile hosts which may generate multi-class traffic. At the base station, an efficient ROS scheme (will be elaborated in Section III) is used as the packet scheduling scheme to allocate bandwidth to mobile hosts. In this system, transmission time is fully slotted. The transmission of packets only starts in the beginning of a time-slot. The scheduler assigns a time-slot to individual users. The users may have different transmission rates, such as that in a hybrid time-division multiple access (TDMA)/CDMA system[8]. The multiple new requests arrived during the time-slot $n-1$ will be accumulated by base station. These requests will be processed by the admission control procedure at the beginning of time-slot n. The sets of existing users and new users are denoted by O and F, respectively. $O = O^v \cup O^d$, where O^v and O^d represent the sets of existing voice users and data users, respectively. In the following, we describe the channel capacity model and the traffic model.

A. Channel Capacity

In wireline networks, the link has fixed capacity, i.e., capacity C = constant. Although in wireless networks the link capacity varies with multiple aspects such as fading, the long-term average capacity ($\bar{C}$) can be viewed as a constant value. Such model is accurate in another regime where the changes of channel states are "covered" by physical layer coding and modulation.

Suppose r^v and r^d represent the maximum transmission rate of voice users and data users, respectively. F_A is the set of newly admitted incoming users. ϕ_i is the pre-assigned queue weight for user i. After the new users are admitted, the allocated bandwidth for user i will be[2]

$$r_i = \frac{\phi_i}{\sum\limits_{j\in O^v}\phi_j + \sum\limits_{k\in O^d}\phi_k + \sum\limits_{l\in F_A}\phi_l}\bar{C}. \tag{1}$$

If $r_i \geq r^v$ for voice users and $r_i \geq r^d$ for data users, the bandwidth requirement is satisfied and no backlog is generated. However, the link capacity is partially wasted and under-utilized in such case. Suppose this link capacity is greater than the transmission requirement of M^v voice users and M^d data users at the same time, i.e., $C \geq M^v r^v + M^d r^d$. Since the probability of all users transmitting at the same time is very small, this system may accommodate more than $M^v + M^d$ users, if temporary congestion is allowed. Since there will exist such a case that $r_i < r^v$ or $r_i < r^d$, the backlog may increase under such condition. Assume each user is allocated buffer with fixed size b_i (for simplicity $b_i = b,\ \forall i$). The buffer will overflow if too many users are admitted to use the wireless link.

B. Traffic Model

On-off traffic model is widely used to describe the burst characteristics of voice and multimedia traffic[6], [8]. In this paper, we use an on-off model with two states—G (on) and B (off). The transition probability from G state to B state is b, while the transition probability from B to G is g. The probabilities that the model remains at G and B states are $1-b$ and $1-g$, respectively. All packets are assumed to have the same packet length.

In the system considered in this paper, we assume the transmission state keeps unchanged during a time slot. The state transition only occurs at the beginning of a time slot. Both the "on" period and the "off" period are assumed to be geometrically distributed, and the average value of these two periods are represented by L_{on} and L_{off}, respectively. Since the "on" state transits to "off" state with probability b, and remains at "on" state with probability $1-b$. Then the probability of a burst contains k time slots is given by

$$P(\text{burst} = k) = b(1-b)^{k-1}. \tag{2}$$

Thus, we have $L_{\text{on}} = 1/b$. Supposing the "off" state remains at current state with probability $1-g$ and transits to another state with probability g, we have the probability of a "off" period containing k time slots as follows,

$$P(\text{period} = k) = g(1-g)^k. \tag{3}$$

Similarly, we have $L_{\text{off}} = \frac{g}{1-g}$. Thus the traffic load of this burst traffic flow is given by

$$\rho = \frac{L_{\text{on}}}{L_{\text{on}} + L_{\text{off}}} = \frac{g}{g + b - bg}. \tag{4}$$

C. Channel Model

In this paper, the two-state channel model[9] is used to describe wireless channel states. In the "good" state, the number of packets can be transmitted per time-slot is R. Otherwise, no packet can be transmitted in "bad" state.

III. RATE GUARANTEED SCHEDULING (ROS)

In this section, we provide an overview of an efficient ROS scheme for wireless networks[2]. The design goal of ROS is to achieve maximum system throughput while satisfying the fairness constraint of resource allocation with a low implementation complexity. Due to the time varying nature of wireless channel, the maximum transmission rates of each user also change over time. Naturally dispatching the user with maximum transmission rate at all times will maximize system throughput. However, the tradeoff between individual user satisfaction and system throughput is necessary. The users with long distance to the base-station or high disturbance from other signal sources will be sacrificed with small transmission opportunity. While high transmission power and transmission opportunity is allocated to those users near the base-station.

Our scheme is motivated by the work of[1], in which the authors present a solution to maximize system performance value by exploiting channel states while satisfying time-fraction assignments. However, this scheme has a high computational complexity in evaluating parameters. The authors also do not provide analysis of performance bound in the paper. Besides work in[1], there are some studies[8], [10] extending scheduling algorithms for wireline networks to wireless networks. They provide short-term or long-term throughput bound for incoming flows. At the same time, they consider the fairness constraint between flows and are able to guarantee some short-term or long-term fairness. In order to satisfy the "fairness" constraint, we employ the measure "relative throughput," which is defined as the throughput divided by the predefined weight. In more details, the difference of "relative throughput" between users is used as a key parameter in the operation of ROS. The greater this parameter, the more the unfairness service received by users.

ROS scheme chooses the flow with highest priority to transmit at the beginning of each time slot. The priority of flow i is defined as

$$\text{prio}_i = R_i + x_i \tag{5}$$

where R_i is the maximum transmission rate of flow i at the next time-slot, and

$$x_i = \max_j\{y_j(t)\} - y_i,\ i, j = 1, 2, \cdots, N$$

where N is the number of flows serviced by this node. $y_i(t)$ is the "relative throughput" which is defined as

$$y_i(t) = O_i(t)/\phi_i \tag{6}$$

where $O_i(t)$ represents the amount of data of flow i which has been serviced, ϕ_i is the normalized rate-fraction of flow i, i.e., $\phi_i = w_i/w_0$, $w_0 = \min\{w_i\}$, where w_i is the predefined queue weight for flow i.

Suppose the system is a concatenation of M nodes, and flow i is regulated by leaky bucket (σ_i, ρ_i) at the ingress node. ROS scheme can provide guaranteed service rate r_i for each flow. The service curve can be described as $\beta_{r_i,v_i} = r_i t - r_i v_i$, where $v_i = \frac{F_i}{r_i}$ and $F_i = N D_{\max}\phi_i / \sum_j \phi_j$. $D_{\max}$ is the upper bound of x_i. The detailed deduction of the upper bound is given in[2]. The service curve provided by an ROS node is given by

$$\beta(t) = s_m t - \sum F_i = s_m t - D_{\max} N$$

where s_m is the bandwidth of output link of node m. According to the delay bound analysis for concatenation of guarantee rate (GR) nodes[11], the delay bound of ROS is

$$d_{\text{bound}} = \sum_{m=1}^{M} \frac{D_{\max} N}{s_m} + L_{\max} \sum_{m=1}^{M-1} \frac{1}{s_m} + \frac{\sum_i \sigma_i}{\min_m s_m} \tag{7}$$

where $L_{\max}$ is the maximum packet length.

By measuring the "relative throughput" difference between users, ROS can provide the mobile hosts rate-guaranteed service, which is similar as the GR node with loss in the model of service curve with loss. ROS chooses the user with maximum "opportunity" to transmit in one time-slot. We prove that ROS can achieve maximum throughput compared with other GR nodes via theoretical analysis and validate the results via simulation. As a GR node with loss, ROS can also provide delay bound for regulated incoming flows. Besides all these advantages, ROS has another important merit—low implementation complexity.

However, without the aid of admission control policy, the number of flows transmitting in the network could be out of control. When there are too many flows sending packets at the same time, there will be a disaster for network and the QoS requirements of all flows may be violated. Therefore, a coordinated admission control policy is necessary for ROS to provide statistical QoS guarantee. In next section, we will introduce our new admission control policy.

IV. NEW ADMISSION CONTROL POLICY—ROS-CAC

In this section, we first derive the probability distribution function of the queue length of individual flows at base station and estimate the PLRs of all flows. PLR is then used as the main admission criteria for the following considerations. PLR itself is an important QoS parameter for realtime multimedia traffic.

In addition, PLR is closely related with other QoS parameters. When congestion occurs in a network, longer packet delay leads to more packet losses. There is a tradeoff space between delay requirement and loss probability of a flow, both of which are affected by the buffer size allocated to the flow. When a new flow is admitted, a certain amount of buffer is allocated to it. The packets belonging to this flow are queued in the buffer and the head of line (HOL) packet will be transmitted if the flow is chosen for service by the ROS scheduler. The smaller the buffer size, the lower the queueing delay at this node because the packets which have been waiting for too long will be discarded. In other words, smaller buffer size leads to a shorter queueing delay but a higher packet loss probability. The optimal value of buffer size should satisfy both delay and packet loss constraints. Throughput is also related with packet losses. Obviously, more packet losses lead to lower throughput. In this work, we focus on the packet loss due to buffer overflow at the scheduler. Then our admission problem is changed to estimating the probability of buffer overflow p_i^o. The admission policy can be expressed as

$$p_i^o \le p_i^d, \quad i \in O \cup F \tag{8}$$

where p_i^d is the PLR requirement for user i, O and F are sets of active and new users, respectively. Before admitting a new user, the base station computes $p_i^o, i \in F$ for new users and re-estimates $p_i^o, i \in O$ for all active users. If all the updated p_i^o satisfy the requirements, this new user is admitted.

A. Analysis of PLR

In this subsection, we derive the PLR caused by buffer overflow in ROS. At each time-slot only one user is permitted to transmit its data. Since each user has its own queue, we can investigate the behaviors of one of these queues without loss of generality. Let l be the index of user who is allocated this queue. Let u_i be the probability that i packets are waiting in this queue, we have

$$\begin{cases} u_0 &= A_0 \sum_{i=0}^{R} u_i \Pr\{d=R\} + pA_0u_0 \\ u_1 &= A_0u_{R+1}\Pr\{d=R\} + A_1 \sum_{i=0}^{R} u_i \Pr\{d=R\} \\ &\quad + p(A_0u_1 + A_1u_0) \\ &\vdots \\ u_n &= A_n \sum_{i=0}^{R} u_i \Pr\{d=R\} + u_0A_np \\ &\quad + \sum_{j=R+1}^{R+n} A_{n+R-j}(u_j \Pr\{d=R\} + u_{j-R}p) \end{cases} \tag{9}$$

where $A_i = \Pr\{i \text{ arrivals in a time-slot}\}$. Here, we suppose that the packet arrival rate follows Poisson distribution. p is the probability of user who is not scheduled to transmit, and $\Pr\{d=R\}$ is the probability that maximum possible number of departure packets ("d") is equal to R, i.e.,

$$\begin{aligned} \Pr\{d=R\} &= \Pr\{\text{user } l \text{ is scheduled to transmit}\} \\ &\times \Pr\{\text{user } l \text{ is in "good" state at the same time}\}. \end{aligned} \tag{10}$$

In[2], we have proved that $\Pr\{\text{user } l \text{ is scheduled to transmit}\} = \frac{\phi_l}{\sum \phi_j}$. The probability of channel staying at "bad" state is small as compared at "good" state. Furthermore, we proved that the probability of scheduled user experiencing "bad" state is much more smaller than that of "good" state[2]. Thus, the second term of (10), $\Pr\{\text{user } l \text{ is in "good" state at the same time}\}$ can be ignored, that is $\Pr\{\text{user } l \text{ is in "good" state at the same time}\} \approx 1$. We have

$$\Pr\{d=R\} = \frac{\phi_l}{\sum_j \phi_j} = 1-p, \ \forall l. \tag{11}$$

Equation (9) can be solved by the generating function $P(z) \equiv \sum_{n=0}^{\infty} u_n z^n$. Defining $Q_R = \sum_{i=0}^{R} u_i$, we have

$$\begin{aligned} P(z) &= [(1-p)Q_R + u_0p] \sum_{n=0}^{\infty} e^{-\rho} \frac{\rho^n}{n!} z^n \\ &\quad + \sum_{n=1}^{\infty} z^n \sum_{i=1}^{n} e^{-\rho} \frac{\rho^{n-i}}{(n-i)!} [(1-p)u_{i+R} + u_ip] \\ &= [(1-p)Q_R + u_0p]e^{\rho(z-1)} \\ &\quad + \sum_{i=1}^{\infty} [(1-p)u_{i+R} + u_ip] z^i \sum_{n=i}^{\infty} e^{-\rho} \frac{\rho^{n-i}}{(n-i)!} z^{n-i} \\ &= [(1-p)Q_R + u_0p]e^{\rho(z-1)} \\ &\quad + e^{\rho(z-1)} \sum_{i=1}^{\infty} [(1-p)u_{i+R} + u_ip] z^i \end{aligned} \tag{12}$$

implying

$$\begin{aligned} P(z) &= \frac{(1-p)Q_Rz^R - (1-p)\sum_{k=0}^{R} u_kz_k}{z^Re^{\rho(1-z)} - (1-p+pz^R)} \\ &= \frac{(1-p)\sum_{k=0}^{R-1} u_k(z^R - z^k)}{z^Re^{\rho(1-z)} - (1-p+pz^R)} \end{aligned} \tag{13}$$

where ρ is the mean packet arrival rate (number of packets arriving per time-slot).

It can be known that the poles of (13) are distinct[12]. According to Rouché's theorem[12], there are $R-1$ zeros of the numerator inside $|z|=1$, and the R-th is clearly $z=1$. Thus, the numerator can be written as

$$N(z) = K(z-1)(z-z_1)\cdots(z-z_{R-1}). \tag{14}$$

By using $P(1)=1$ and L'Hôpital's rule

$$\begin{aligned} 1 = \lim_{z\to 1} P(z) &= \frac{K(1-z_1)\cdots(1-z_{R-1})}{(1-p)R - \rho} \\ \Rightarrow K &= \frac{(1-p)R-\rho}{(1-z_1)\cdots(1-z_{R-1})}. \end{aligned} \tag{15}$$

Then,

$$\begin{aligned} P(z) &= \frac{(1-p)R-\rho}{(1-z_1)\cdots(1-z_{R-1})} \\ &\quad \times \frac{(z-1)(z-z_1)\cdots(z-z_{R-1})}{z^Re^{\rho(1-z)} - (1-p+pz^R)}. \end{aligned} \tag{16}$$

The roots $\{z_i, i = 1, \cdots, R-1\}$ can be derived from $z^R = \frac{1-p}{e^{\rho(1-z)}-p}$. By evaluating (16) at $z = 0$, we have

$$u_0 = \frac{[(1-p)R - \rho](-1)^{R-1} \prod_{i=1}^{R-1} z_i}{(1-p) \prod_{i=1}^{R-1} (1 - z_i)}, \ (R \geq 2) \tag{17}$$

$\{u_i, \ i = 1, \cdots, R-1\}$ can be derived from a $(R-1) \times (R-1)$ complex-valued linear system of equations. These equations are created by setting the numerator of (13) to be equal to zero at $R-1$ roots $\{z_i\}$. Then, $\{u_i, \ i \geq R\}$ can be derived from recursion on (9).

Next, we consider the case of finite buffer. The packet loss probability of finite buffer can be estimated by calculating the sum of state probabilities $\sum_{i=B+1}^{\infty} u_i$ (supposing B is the buffer size) for the infinite buffer case as in[13].

B. Admission Control Policy

In this subsection, we deduce the packet loss probabilities of each active flow and the new admitted flow based on the results in previous subsection, and thus design the admission policy. The traffic load (in terms of mean packet arrival rate) ρ can be derived from the traffic model of each flow. For on-off traffic model, the equivalent traffic load can be deduced from (4). Then for each flow l, u_n^l, the probability of n packets staying in the buffer of flow l can be derived. Furthermore, we have the packet loss rate for flow l.

$$\text{PLR}_l(B) = 1 - \sum_{i=0}^{B} u_i^l \tag{18}$$

where B is the buffer size (in number of packets) for flow l.

By using the PLR result from (18), we can design an efficient admission control policy by using the PLRs as the criteria. The admission region is defined in (8). If the calculated $\text{PLR}_l(B)$ is smaller than the PLR requirement of flow l and all existing flows' PLRs are smaller than their requirements, flow l can be admitted. Otherwise, flow l is rejected.

With the increase of number of flows and types of traffic, the computation complexity of queue length PDF and the corresponding PLR will increase accordingly. An near-optimal but simpler algorithm could be developed and this could be an interesting research topic in the future. Another possible solution to reduce the complexity is to make computation "off-line"—before the operation of system, if the parameters of traffic and channel capacity are known. By calculating the results of all possible combinations of traffic flows, the admission region can be determined. The results can be stored in a table, which can be looked up before an admission decision is made.

C. Example of Admission Control Operation

Here, we use a simple example to illustrate the operation of ROS-CAC. We assume that there are N users, including active and newly admitted users, in the system. All users have the same queue weight, i.e., $\phi_i = \phi_j, \forall i, j \in 1, \cdots, N$. Thus all users have the same probability to be scheduled to transmit, i.e., $1 - p = 1/N$. We assume that the maximum number of packets can be transmitted in a time slot $R = 3$. All flows follow the on-off traffic model, and the traffic load ρ of each flow is normalized in terms of number of packets arriving during a time-slot.

At first, we resolve roots $\{z_1, z_2\}$ (obviously $z_3 = 1$) of the following equation

$$z^3 = \frac{1-p}{e^{\rho(1-z)} - p}. \tag{19}$$

The results will be a pair of complex value. Suppose

$$z_1 = x + jy, \ z_2 = x - jy. \tag{20}$$

Thus, according to (17), we have

$$u_0 = \frac{(3(1-p) - \rho)(x^2 + y^2)}{(1-p)((1-x)^2 + y^2)}. \tag{21}$$

From the 2×2 complex-valued linear system of equations created from the numerator of (13), we have

$$\begin{cases} u_1 = u_0 \dfrac{x^2 + y^2 + 2x}{-x^2 - y^2} \\ u_2 = u_0 \dfrac{1 + 2x}{x^2 + y^2}. \end{cases} \tag{22}$$

$\{u_i, i \geq 3\}$ can be derived from recursion of (9).

$$\begin{cases} u_3 = \dfrac{(1 - pA_0)u_0 - A_0 \sum_{i=0}^{2} u_i}{A_0(1-p)} \\ \vdots \\ u_{3+n} = \dfrac{1}{A_0(p-1)} \{A_n \sum_{i=0}^{3} u_i(1-p) + u_n(A_0 p - 1) \\ \qquad + \sum_{j=4}^{2+n} A_{n+3-j}(u_j(1-p) + u_{j-3}p) + u_0 A_n p\}. \end{cases} \tag{23}$$

Next, we show an example of the calculation result of previous procedure. In this example, $N = 20, R = 3$. By setting traffic load ρ to three different values, the corresponding three PDF curves are shown in Fig. 1. Obviously, the lower the traffic load ρ, the higher the probability queue length being short.

The PDF results can be used in admission control policy. Suppose the PLR requirement of the flow is no greater than 0.01, traffic load of each flow is $\rho_i = 0.02$ and queue length allocated to each flow is 20 packets. From the obtained PDF results, we can find that when total traffic load $\rho = 0.10$, PLR $= 0.00922$ the PLR requirements of all flows are satisfied. However, when $\rho = 0.12$, PLR $= 0.096$ the PLR requirements are violated. Thus, the maximum number of users who can be admitted is 5.

D. Discussions

When a new incoming flow is admitted by the admission control policy, the base station allocates a certain size of buffer to it. The packets of this flow are served on the fist come first serve

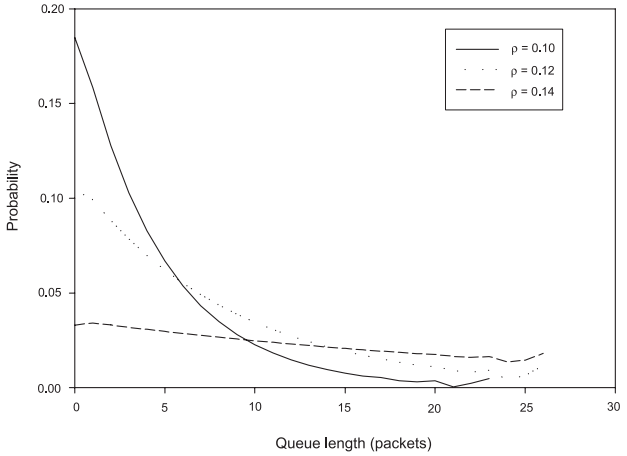

Fig. 1. Probability distribution of queue length in ROS for $\rho = 0.1, 0.12,$ and 0.14.

(FCFS) basis in this buffer. The HOL packet will be transmitted when the flow obtains transmission opportunity. The buffer size B is an important parameter for QoS provisioning. First, PLR of each flow is determined by B. Second, it can also affect the waiting time of packets. The maximum waiting time is $D_m = \frac{B}{r_i}$ in a simplified case. Thus, we have the maximum buffer size not violating delay requirement $B_d = r_i D_i^{\max}$, where $D_i^{\max}$ is the maximum queueing delay requirement of flow i. In short, the buffer management is very important. Supposing the minimum buffer size satisfying the PLR requirement is B_p, the actual allocated buffer size should satisfy following conditions.

$$B \geq B_p \text{ and } B \leq B_d. \tag{24}$$

If B_d is larger than B_p, B can be assigned any value between them.

The computation of maximum waiting time $D_i = \frac{B_d}{r_i}$ is valid only in ideal continuous fluid model. In ROS, the distribution of waiting time is more complicated since only one user can transmit in a time-slot and no simultaneous transmission is allowed. In long-term average, user i is scheduled to transmit with a specified probability. Another particular characteristic of ROS is that packets leave the buffer in a batch during a time-slot. One batch of packets consists of R packets. When a packet enters the buffer of user i, there are already M_n packets waiting in the buffer. The average waiting time (in terms of time slot) of this packet will be

$$\bar{D}_i = (M_n \bmod R)\bar{T}_i + 1 \tag{25}$$

where $\bar{T}_i$ is the average waiting time of HOL packet of this buffer. Since the probability of user i being chosen to transmit is $p_i = \frac{\phi_i}{\sum \phi_j}$, $\bar{T}_i$ can be deduced as follows.

$$\bar{T}_i = 1/p_i = \frac{\sum \phi_j}{\phi_i}. \tag{26}$$

Instead of fixed buffer size, variable buffer size should be more efficient. For example, in the case of B_d larger than B_p, we have $B_p \leq B \leq B_d$. If B is assigned the minimum value B_p, this flow occupies minimum buffer resource. However, some packets may be prematurely dropped before their delay timer expired in such case. Thus packet loss rate increases. On the other side, if B is assigned the maximum value B_d, the PLR will be smaller with more occupied buffer resource. Thus, there is a tradeoff between PLR and buffer resource. The optimal size is an interesting issue to study.

Table 1. Simulation parameters.

Parameter	Value
t_{on}	180 ms
t_{off}	60 ms
r_{on}	128 kB/s
t_{active}	180 s

V. PERFORMANCE EVALUATION

In this section, we use experiments to validate the effectiveness of the proposed joint packet scheduling/admission control framework based on ns-2[7] simulations. In the experiments, we consider a system with a center cell surrounded by six cells placed hexagonally. The performance results of the center cell will be simulated and analyzed. We assume on-off video traffic flows are transmitting in this system. The on-off traffic model generates packets in a constant rate r_{on} during "on" time periods, while not generating during "off" time periods. The lengths of "on" and "off" time periods follow exponential distribution with mean values of t_{on} and t_{off}, respectively. The mean transmission time of each video flow is set to t_{active}. We use the two-state Markov model to describe channel states. In the "good" state, the maximum number of packets can be transmitted in a time-slot will be $R = 3$. Users arrive at the center cell following Poisson distribution. The queue length for each admitted user is fixed at 20 packets. All packets have the same packet length—256 kbytes. Suppose the PLR requirement of the flow is 0.01. Other simulation parameters are listed in Table 1.

A. Video Traffic Scenario

We simulate the new proposed admission control policy (ROS-CAC) based on the traffic model and channel model mentioned previously. We compare the system performance results in two different conditions: Without admission control and with admission control.

The comparison of average packet loss ratio with admission control (AC) and without AC is shown in Fig. 2. When the traffic load is below 1.67/min, the PLR difference between two cases is invisibly small. The PLR requirements are satisfied under both cases. However, the PLR of the system without AC increases very fast with the traffic load, and the PLR requirement is violated when the user arrival rate is greater than 2.5/min. As a comparison, the PLR of the system with AC remains stable and PLR requirement is always guaranteed to be smaller than 0.01. Obviously, the base station can avoid lots of packet losses with the aid of admission control. Moreover, the average end-to-end delay of packets is significantly reduced, which is illustrated in Fig. 3. This is because that the number of packets waiting in queues with admission control is much less than that without

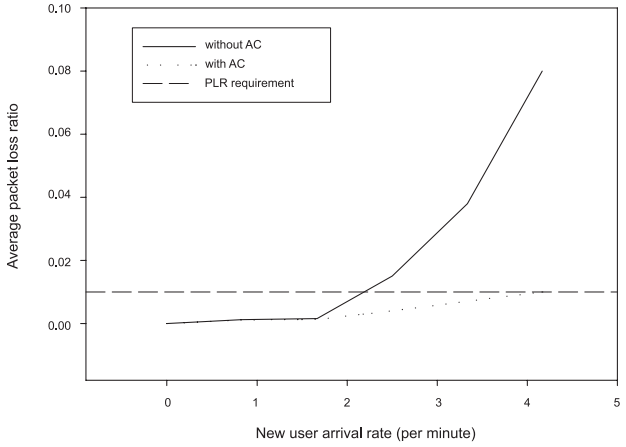

Fig. 2. Comparison of packet loss rate with and without admission control.

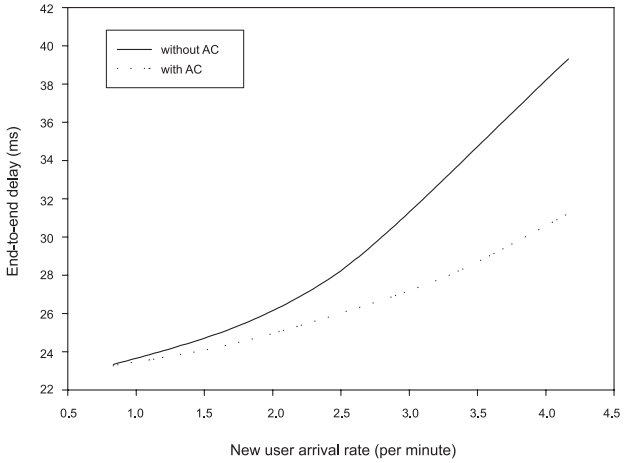

Fig. 3. Comparison of average end-to-end delay with and without admission control.

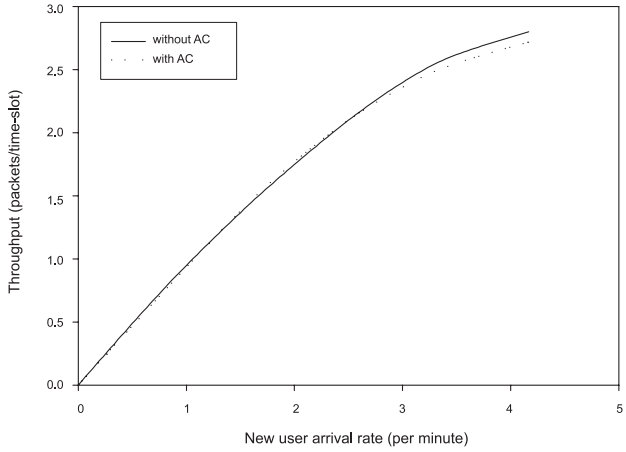

Fig. 4. Comparison of system throughput with and without admission control.

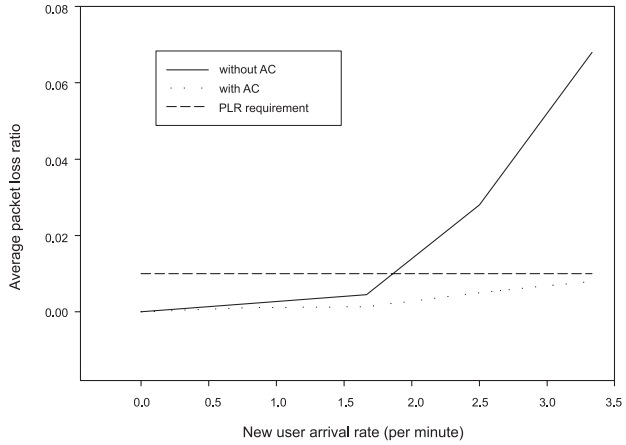

Fig. 5. Comparison of packet loss rate with and without admission control.

admission control.

We evaluate the system throughput by calculating the average number of packets successfully transmitted per time-slot. The simulation results of system throughput are demonstrated in Fig. 4. We can see that the system throughput with admission control is slightly lower than that without AC when new user arrival rate is high. The main reason is that with admission control, less users can transmit simultaneously than that without AC.

B. Mixed Traffic Scenario

Different from previous simulation experiment with only video traffic flows, in this experiment we add two TCP flows in the simulation. Suppose the queue weight of TCP flow $\phi_t = \frac{1}{2}\phi_v$, where ϕ_v is the queue weight of video flow. The traffic model for video flow is the same as that in previous section. We also compare the system performance results between with AC and without AC.

Fig. 5 shows the comparison of average packet loss ratio with admission control (AC) and without AC in mixed traffic scenario. When the traffic load is below 0.75/min, the PLR difference between two cases is invisibly small. The PLR requirements are satisfied under both cases. However, the average PLR of the system without AC increases very fast with the traffic load, and the PLR requirement is violated when the user arrival rate is greater than 1.86/min. As a comparison, the average PLR of the system with AC remains stable and PLR requirement is always guaranteed to be smaller than 0.01. In Fig. 6, we illustrate the average end-to-end delay of packets. We can find that the delay is significantly reduced due to smaller waiting queue length. The system throughput is still evaluated by calculating the average number of packets successfully transmitted per time-slot. The simulation results of system throughput are shown in Fig. 7. Similar as the video traffic scenario, we can see that the system throughput with admission control is slightly lower than that without AC when new user arrival rate is high.

VI. CONCLUSION

In this paper, we first review a novel packet scheduling scheme we proposed for wireless networks, called ROS. In order to provide statistical QoS guarantees, we have proposed a new admission control policy to work together with ROS for wireless networks. Through mathematical analysis, we investigated the behavior of buffer at base-station and packet overflow

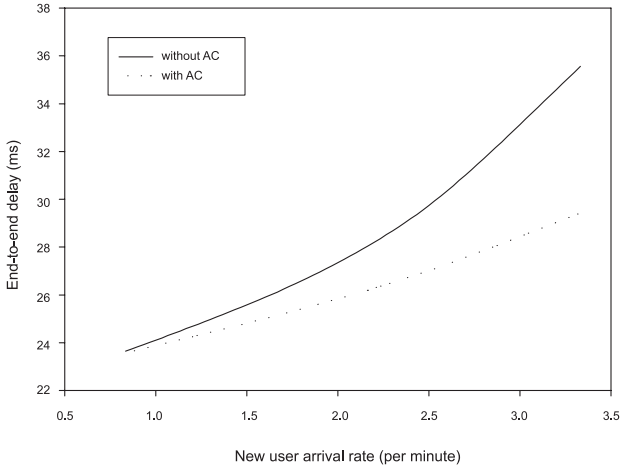

Fig. 6. Comparison of average end-to-end delay with and without admission control.

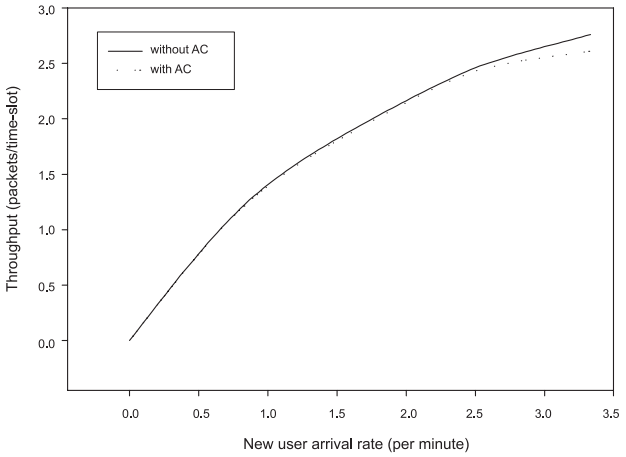

Fig. 7. Comparison of system throughput with and without admission control.

probability. In order to maintain low packet loss rates for mobile users, the base-station allocates buffer of an appropriate size to individual users. Our proposed new admission control policy admits new user if all active users' PLR will not be violated. The simulation results have validated the effectiveness of our proposal.

REFERENCES

[1] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, pp. 2053–2064, Oct. 2001.

[2] F. Long, G. Feng, and C. K. Siew, "Efficient rate guaranteed opportunistic scheduling," *IEEE Trans. Mobile Computing*, submitted for publication, available at http://www.ntu.edu.sg/home5/pg03753229/lftmc.pdf.

[3] P. Das and J. Y. Khan, "Quality of service based resource management for packet switched data over the WCDMA uplink," in *Proc. IEEE ICC 2004*, June 2004.

[4] M. Hu and J. Zhang, "Opportunistic multi-access: Multiuser diversity, relay-aided opportunistic scheduling, and traffic-aided smooth admission control," *Mobile Networks and Applications*, vol. 9, pp. 435–444, 2004.

[5] K. Mase, Y. Toyama, A. A. Bilhaj, and Y. Suda, "QoS management for VoIP networks with edge-to-edge admission control," in *Proc. IEEE GLOBECOM 2001*, vol. 4, Nov. 2001, pp. 2556–2560.

[6] T.-K. Liu and J. A. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services," *IEEE J. Select. Areas Commun.*, vol. 16, no. 6, pp. 845–857, Aug. 1998.

[7] "The Network Simulator—ns-2," available at http://www.isi.edu/nsnam/ns/index.html.

[8] F. Long, G. Feng, and C. K. Siew, "Channel states dependent fair service: A new packet scheduling algorithm for CDMA," *Computer Networks*, vol. 49, no. 2, pp. 201–216, 2005.

[9] L. Wilhelmsson, "On the effect of imperfect interleaving for the Gilber-Elliott channel," *IEEE Trans. Commun.*, vol. 47, no. 5, pp. 681–688, May 1999.

[10] T. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," in *Proc. IEEE INFOCOM'98*, vol. 3, 1998, pp. 1108–1111.

[11] J.-Y. Le Boudec and P. Thiran, *Network Calculus, A Theory of Deterministic Queuing Systems for the Internet*, Springer Verlag LNCS 2050, 2004.

[12] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Chapter 6, John Wiley & Sons, 1998.

[13] J. M. Pitts and J. A. Schormans, *Introduction to ATM Design and Performance*, John Wiley & Sons, 1997.

[14] V. G. Kulkarni, L. Gun, and P. F. Chimento, "Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1039–1047, Aug. 1995.

[15] S. Khatun, B. M. Ali, V. Prakash, and M. Ismail, "Performance analysis and optimization of a mobility support ATM switch," in *Proc. GLOBECOM 2001*, Nov. 2001, vol. 6, pp. 3449–3453.

[16] H. S. Kim and N. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers," *IEEE Trans. Networking*, vol. 9, pp. 755–767, Dec. 2001.

[17] J. Choe and N. B. Shroff, "Queueing analysis of high-speed multiplexers including long-range dependent arrival processes," in *Proc. IEEE INFOCOM'99*, Mar. 1999, vol. 2, pp. 617–624.

[18] Y.-K. Kwok and V. K. N. Lau, "On admission control and scheduling of multimedia burst data for CDMA systems," *Wireless Networks*, pp. 495–506, Aug. 2002.

[19] X. Wang, "Wide-band TD-CDMA MAC with minimum-power allocation and rate- and BER-scheduling for wireless multimedia networks," *IEEE Trans Networking*, vol. 12, pp. 103–116, Feb. 2004.

[20] Y. Cheng and W. Zhuang, "Calculation of loss probability in a partitioned buffer with self-similar input traffic," in *Proc. IEEE GLOBECOM 2004*, 2004, pp. 1453–1457.

[21] Z. Liu, M. J. Karol, M. E. Zarki, and K. Y. Eng, "Channel access and interference issues in multi-code DS-CDMA wireless packet (ATM) networks," *Wireless Networks*, vol. 2, pp. 173–193, 1996.

[22] H. Varian, "Equity, envy, and efficiency," *J. Economic Theory*, vol. 9, pp. 63–91, 1974.

[23] R. Mazumdar, L. Mason, and C. Douligeris, "Fairness in network optimal flow control: Optimality of product forms," *IEEE Trans. Commun.*, vol. 39, pp.775–782, May 1991.

[24] P. Bender and P. Black, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, pp. 70–77, Jul. 2000.

Fei Long received his B.Eng. and M.Eng degrees in Electronic Engineering from the Tsinghua University in 1999 and 2002, respectively. He entered the Information Communication Institute of Singapore, Nanyang Technological University in March 2003 to pursue Ph.D. degree. His research mainly focus on traffic control for wireless networks.

Gang Feng received his B.Eng. and M.Eng. degrees in Electronic Engineering from the University of Electronic Science and Technology of China, in 1986 and 1989, respectively, and the Ph.D. degree in Information Engineering from The Chinese University of Hong Kong in 1998. He joined the Information Communication Institute of Singapore, Nanyang Technological University in August 1999. Before that, he worked for about one year in the Department of Electronic Engineering, City University of Hong Kong as a postdoc. From 1989 to 1995, he was with the Research Institute of Information Systems, University of Electronic Science and Technology of China. He is currently an associate professor in the School of EEE, Nanyang Technological University. Dr. Feng's research interests include routing and performance evaluation for high-speed networks, TCP enhancement over heterogeneous networks. Recently, he branches out to work on reliable multicast, flow and congestion control in the Internet, and admission control and QoS for multiservice wireless networks.

Junhua Tang received her B.Eng. and Ph.D. degrees in Electrical and Electronic Engineering from Shanghai Jiao Tong University (China) in 1994 and 1999, respectively. She worked in the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, as a research fellow from April 1999 to November 2000. From November 2000 to December 2001, she worked with Lucent Technologies Singapore Pte Ltd. as a technical consultant. She joined NTU in Jan 2002 as a teaching fellow and is an assistant professor since April 2005. Her research interests include QoS in both wired and wireless networks, network traffic engineering, and network performance analysis.