This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2007 proceedings.

Losing Opportunism: Evaluating Service Integration in an Opportunistic Wireless System*

Hongseok Kim and Gustavo de Veciana
Wireless Networking and Communications Group (WNCG)
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
{hkim4, gustavo}@ece.utexas.edu

***Abstract*— In this paper we evaluate interactions among flow-level performance metrics when integrating QoS and best effort flows in a wireless system using opportunistic scheduling. We introduce a simple flow-level model capturing the salient features of bandwidth sharing for an opportunistic scheduler which ensures a mean throughput to each QoS stream for every time slot. We then explore the flow-level performance showing that integration of QoS and best effort flows results in *loss in opportunism*, which in turn results in a reduction of the stability region, degradation in system throughput, and increased file transfer delay. These losses are shown to be proportional to opportunistic gains, the guaranteed bandwidth and number of QoS flows, but inversely proportional to SNR under a Rayleigh fading channel model. In an integrated system exploiting opportunism, local instability appears to be more severe than in wired networks and average delays experienced by best effort flows are prolonged. We suggest that a form of admission control for best effort flows is necessary to avoid local instability, and ensure adequate performance.**

I. INTRODUCTION

Wireless networks are evolving towards supporting multiple services, *e.g.,* both best effort and QoS streaming traffic. Since the integration of different services on a single platform is expected to generate new revenue and reduce network management cost, intensive research efforts have been devoted towards designing such networks. However, because wireless resources are limited and shared by users experiencing time-varying channels, service integration may be quite challenging. A key element in such systems is the traffic scheduler and complementary resource management component that can assure users appropriate QoS. In addition, such schedulers may be designed to be opportunistic, *i.e.,* serve users whose current channel capacities are high. Attempting to be opportunistic while meeting users' QoS requirements presents significant new challenges, see [1]–[11].

Opportunistic scheduling schemes developed so far are mostly packet-level algorithms [1]–[4], [6], [7], [9]–[11] focusing on the case where the user population is static and queues are infinitely backlogged. This assumption is meaningful in short time scales where the user population does not change much. Under this assumption, Liu *et al.* propose throughput optimal solutions under temporal and utilitarian fairness criteria [3], [4]. Patil *et al.* and Park *et al.* independently propose a scheduling scheme based on a history of channel information [7]–[9]. However, static approaches may not capture the flow-level dynamics in which new flows come to the system randomly and leave after being served. In a real system, user population changes over time and it is of interest to know system performance such as average throughput and average file transfer delays. This problem was first addressed by Borst who proposed flow-level analysis using multi-class processor sharing model [5]. However, this work did not deal with the mixes of QoS and best effort traffic. Other attempts to integrate QoS and best effort in wireless opportunistic systems have been recently done by [6], [8], [10]. But, these studies focused on packet-level performance only.

Hence, we are motivated to study a new model that addresses the interaction of heterogeneous traffic at the flow-level. In *wired* network case, there had been several studies on this topic [12]–[19]. To analyze the interaction of elastic and streaming flows, researchers have used a 2-dimensional Markovian model [15]–[17]. The work of Key *et al.* suggests that the integration of heterogeneous traffic has a positive consequence, *i.e.,* stabilizing effect [15]. The work in [13] highlights that such systems are likely to see transient, or local, instability. Specifically when there are too many QoS sessions, the best effort flows may accumulate, but subsequently subside once more bandwidth becomes available, *i.e.,* QoS sessions leave the system. In [14], [19] the authors propose an integrated admission control for both of streaming and elastic traffic to guarantee QoS. Note that these studies were done at a higher level, *i.e.,* considering the integration of TCP and UDP and where the service rate of wired network is constant. The major difference that arises in wireless networks, however, lies in that the service rate of a wireless network is time-varying, and, furthermore, may be shared in an opportunistic manner making the analysis somewhat challenging. The main goal of this paper is to model and study the flow-level characteristics for an opportunistic wireless system shared by a traffic mix of QoS streaming and best effort traffic.

To this end, we propose a new flow-level model for such a system. The scheduler is designed to guarantee a mean

* This work was partially supported by a grant from Freescale.

0743-166X/07/$25.00 ©2007 IEEE

throughput to streaming media irrespective of the number of ongoing flows by borrowing/lending bandwidth from/to best effort flows. Thus, QoS flows have a fixed average throughput per slot which might be set to roughly meet their QoS requirements. Other QoS metrics such as delay or jitter are not considered here to keep the model simple. By contrast, the performance metric for best effort flows will be the *average delay* to finish a finite size of file transfer using HTTP or FTP. Our QoS definition is simple but it gives us an insight of evaluating service integration.

Contributions: The following are the key contributions of this paper.

1) To our knowledge, this paper is the first to attempt to investigate the flow-level interaction between QoS and best effort traffic in systems exploiting opportunism. Our model is simplistic but significant in that it incorporates the essential ingredients of flow-level behaviors such as QoS requirements, opportunistic sharing, stability and evaluation of the stationary distribution. We also identify the necessary and sufficient stability condition of 2-dimensional Markov chain.
2) We show how the integration of QoS and best effort flows compromises the benefits of opportunism in crucial aspects; stability region reduction, system throughput degradation and increased file transfer delay. All of these negative impacts are called *loss in opportunism* of integration. Our analysis shows that, for example, introducing a single QoS user requesting 300kbps would degrade by 33% the maximum system capacity in the CDMA/HDR system described in [1]. We will show that such losses increase in proportion to the opportunistic gains, the number of QoS users and the guaranteed bandwidth, but is inversely proportional to SNR in Rayleigh fading channel model.
3) As in the wired case [13], [17], even if the system is stable, it may exhibit local instability for best effort users. However it appears to be a more crucial phenomenon in wireless systems exploiting opportunism. To circumvent this problem, we suggest that admission control of best effort flows is necessary.

The paper is organized as follows. In Section II we describe our flow-level model for mixed traffic based on bandwidth borrowing and lending among traffic types. We also build up a compact model and analysis tool to investigate flow-level dynamics. Section III is devoted to the stability of the system. Section IV evaluates the opportunistic losses of integration by quantifying the reduction in the stability region, throughput degradation and increase in delay. Section V deals with the performance impact of local instability and the necessity of call admission control, and is followed by conclusions in Section VI.

II. System model

A. Assumptions

We consider a wireless access point shared by multiple mobile users. Wireless access point is assumed to accommodate multiple users by Time Division Multiple Access (TDMA) scheme where time is divided into equal-sized slots and at most one user gets served per slot. We assume that channel capacity for each user is a stationary ergodic process and these processes are independent, identically distributed (*i.i.d.*) across users. This assumption allows us to adopt max rate scheduling as a basic scheduling policy. It maximizes total capacity or throughput of the system [20], [21]. (We will use the term *capacity* or *throughput* interchangeably in the sequel.) The scheduling policy will be revised to model the need to meet QoS session requirements. For simplicity, we divide users into two groups: QoS and best effort users. [1] The channel capacity for each user is independent across slots and remains constant during a slot, *i.e.,* we assume fast fading channel, with a slot time corresponding to coherent time. In the sequel we assume time slots are small relative to flow dynamics and time-scale, and we will model the system dynamics based on a continuous-time model.

A sustained throughput is one of the basic requirements for QoS, so we assume that QoS users are guaranteed a mean throughput $\bar{b}$ per time slot irrespective of the number of best effort users. Our notion of QoS, however, does not guarantee delay constraints. In fact, one might argue that this assumption is not realistic because real-time interactive applications such as voice or video communication will require delay constraints. Nevertheless, if every time slot, QoS users get an average throughput $\bar{b}$, it is very unlikely that QoS user is starved for long period of time slots, which implies decent delay performance. This simple model roughly captures the throughput loss in the system - it is likely to be worse if a more sophisticated scheduling scheme meeting QoS requirements is used. Furthermore, for non real-time one-way streaming media such as video on demand, our QoS notion makes sense because we can assume that end-user devices have buffer space to compensate delay or jitter occurred during transmission.

B. Flow-level model of mixed traffic

In our flow-level model, QoS and best effort flows arrive randomly and leave after being served. We assume that the arrivals of QoS flows follow a Poisson process with arrival rate λ_q and have a holding time which is exponentially distributed with mean μ_q^{-1}. The maximum number of QoS flows is limited to n^* in order to guarantee a bandwidth $\bar{b}$. We also assume that the arrivals of best effort flows follow an independent Poisson process with arrival rate λ_b and have file sizes which are exponentially distributed with mean μ_b^{-1}.

Let $N_q(t)$ be the number of QoS flows and $N_b(t)$ be the number of best effort flows in the system. Then, $(N_q(t), N_b(t))$ is a 2-dimensional Markov process with state space $\{0, \ldots, n^*\} \times \mathbb{Z}^+$. Since our model is an opportunistic system, the available capacity for QoS and best effort flows depends on how they are scheduled. Let $g(n_q, n_b)$ denote the average system capacity given (n_q, n_b). Then, the total

[1] Even under this assumption, each user can still receive both QoS and best effort services using TDMA.

capacity required for QoS flows is $n_q\bar{b}$ and the capacity available to best effort flows is $g_b(n_q, n_b) := g(n_q, n_b) - n_q\bar{b}$. The rate matrix for the chain is then given by:

$$\begin{aligned} q\big((n_q,n_b),(n_q+1,n_b)\big) &= \lambda_q \mathbf{1}_{\{n_q<n^*\}}; \\ q\big((n_q,n_b),(n_q,n_b+1)\big) &= \lambda_b; \\ q\big((n_q,n_b),(n_q,n_b-1)\big) &= g_b(n_q,n_b)\mu_b \mathbf{1}_{\{n_b\geq 1\}}; \\ q\big((n_q,n_b),(n_q-1,n_b)\big) &= n_q\mu_q. \end{aligned} \quad (1)$$

Note that the number of QoS sessions follows an M/M/m/m-like system so the stationary distribution of QoS flows $\pi_q(n_q)$ is independent of n_b and given by

$$\pi_q(n_q) = \pi_q(0)\rho_q^{n_q}\frac{1}{n_q!} \quad (2)$$

where $\rho_q = \frac{\lambda_q}{\mu_q}$ and $\pi_q(0) = \big[\sum_{n_q=0}^{n^*}\rho_q^{n_q}\frac{1}{n_q!}\big]^{-1}$. The blocking probability of QoS flows is given by *Erlang-B formula* as $\pi_q(n^*)$ [22]. Meanwhile the dynamics of the number of best effort flows follow a processor sharing system with varying capacity.

C. Proposed opportunistic scheduling

Suppose that at a given time we have total number of flows n. The total number of flows is the sum of n_q QoS flows and n_b best effort flows. Let X_i, $i \in \{1,\ldots,n\}$ be a random variable representing channel capacity of user i. Since all users are symmetric, the maximum system capacity is given by $g(n) := \mathbb{E}[X^{(n)}]$ where $X^{(n)} \triangleq \max[X_1,\cdots,X_n]$ and is shared equally among the users. So the bandwidth per user $h(n) := \frac{g(n)}{n}$ is decreasing in n while $g(n)$ is increasing in n as shown in Fig. 1. Now, for every time slot, we want to guarantee $\bar{b}$ to QoS flows. If $n \leq n^*$ where

$$n^* := \max\{n | h(n) \geq \bar{b}, n \in \mathbb{Z}^+\},$$

every user has a bandwidth of at least $\bar{b}$ and we satisfy the QoS requirement. We refer to $\{(n_q,n_b)|n_q+n_b \leq n^*\}$ as the *normal regime.*

However, if $n_q + n_b > n^*$, QoS users will not meet their requirement $\bar{b}$. How can we guarantee $\bar{b}$ to QoS flows in the *overloaded regime* $\{(n_q,n_b)|n_q+n_b > n^*\}$? To do this, we propose to use *bandwidth borrowing* as follows. If $h(n)$ is below $\bar{b}$, *i.e.,* $n > n^*$, then QoS flows borrow time slots from best effort flows. Similarly, if $h(n)$ is over $\bar{b}$, then QoS flows lend their time slots to best effort flows. As a consequence, the average throughput of QoS is $\bar{b}$ in every time slot. Under this model we still need admission control for QoS flows to ensure $n_q \leq n^*$. These borrowing and lending mechanisms are described in more detail below.

1) Capacity balance equation in the overloaded regime: The balance equation is given by

$$h(n_q+n_b) + \frac{\alpha(n_q,n_b)}{n_q}\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]P_b = \bar{b} \quad (3)$$

where $P_b := P\big(X^{(n_b)} > X^{(n_q)}\big)$ and $\alpha(n_q,n_b)$ is the *borrowing* probability. The intuition for the equation is as follows. The amount of bandwidth each QoS flow must borrow is $\bar{b} - h(n_q+n_b)$. To compensate this deficiency, we will randomly, with probability $\alpha(n_q,n_b)$, give a best effort slot, *i.e.,* one where $X^{(n_b)} > X^{(n_q)}$, to the QoS user currently seeing the best channel. Thus, the total borrowed bandwidth is $\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]P_b$ and it is shared by n_q QoS flows to meet the guaranteed bandwidth $\bar{b}$.

Note that, in the overload regime, to maintain the average throughput $\bar{b}$, at some time slots, a best effort user which currently has the best channel amongst all users may have to give the time slot to QoS user. Thus, meeting QoS requirements inevitably degrades the overall system throughput. We will study this in detail in Section IV.

2) Capacity balance equation in the normal regime: Assuming QoS users do not need more bandwidth than $\bar{b}$, we reallocate the excess bandwidth of QoS flows to best effort flows, and the capacity balance equation in the normal regime is given by

$$h(n_q+n_b) - \frac{\beta(n_q,n_b)}{n_q}\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}]P_q = \bar{b} \quad (4)$$

where $\beta(n_q,n_b)$ is the *lending* probability and $P_q := P\big(X^{(n_b)} \leq X^{(n_q)}\big)$. [2]

Solving (3) and (4) we can obtain the opportunistic system capacity of $g(n_q,n_b)$. Since QoS users always have average throughput $\bar{b}$, the capacity of best effort flows at (n_q,n_b) state is $g_b(n_q,n_b) = g(n_q,n_b) - n_q\bar{b}$.

D. Capacity of best effort flows in overloaded regime

To solve the capacity balance equation of (3) and (4), we first compute the conditional capacity of QoS users $\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]$. We will do this as follows:

$$\begin{aligned} &\mathbb{E}[X^{(n_q)}|X^{(n_b)} > X^{(n_q)}]P_b \quad + \\ &\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}]P_q \quad = \quad \mathbb{E}[X^{(n_q)}] \end{aligned} \quad (5)$$

TABLE I
NOTATION SUMMARY

n_q	number of QoS flows in a system
n_b	number of best effort flows in a system
n^*	maximum number of QoS flows
$\bar{b}$	average throughput of QoS flows
X_i	a random variable representing channel capacity of user i.
$X^{(n)}$	$\max[X_1,\cdots,X_n]$
$g(n)$	$\mathbb{E}[X^{(n)}]$
$h(n)$	$\frac{g(n)}{n}$
$g(n_q,n_b)$	the system capacity for n_q QoS and n_b best effort flows.
$g_b(n_q,n_b)$	$g(n_q,n_b) - n_q\bar{b}$, the capacity of best effort users.
$h_b(n_q,n_b)$	$\frac{g_b(n_q,n_b)}{n_b}$, the individual capacity of best effort user.
$\bar{g}(n_q)$	$\lim_{n_b\to\infty} g(n_q,n_b)$
$\bar{g}_b(n_q)$	$\lim_{n_b\to\infty} g_b(n_q,n_b)$
$g_b^*(n_q,n_b)$	$g(n_q+n_b) - n_q\bar{b}$
$\alpha(n_q,n_b)$	bandwidth borrowing probability
$\beta(n_q,n_b)$	bandwidth lending probability
$\xi(\bar{b},n_q)$	the capacity gap at $\bar{b}$ and n_q
C	maximum system capacity
κ	$\frac{C}{E[X]}$, opportunistic gain
η	call blocking probability

[2] Since we assume that X_i is a continuous random variable, the equality of $X^{(n_b)} = X^{(n_q)}$ can be placed either in P_b or P_q. In a discrete case which is more likely in the case in practice, we need a tie-breaking rule.

where $P_b = P(X^{(n_b)} > X^{(n_q)}) = \frac{n_b}{n_q+n_b}$ and $P_q = 1 - P_b$. Note that

$$\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}] \quad = \quad \mathbb{E}[X^{(n_q+n_b)}]. \tag{6}$$

Combining (3), (5), (6), we obtain

$$h(n_q + n_b)(1 - \alpha(n_q, n_b)) + h(n_q)\alpha(n_q, n_b) = \bar{b}$$

which means that the system in overloaded regime is identical to one operating as if it had only n_q flows with probability $\alpha(n_q, n_b)$ and one with $n_q + n_b$ flows with probability $1 - \alpha(n_q, n_b)$. So, the borrowing probability is

$$\alpha(n_q, n_b) \quad = \quad \frac{\bar{b} - h(n_q + n_b)}{h(n_q) - h(n_q + n_b)}.$$

The numerator of $\alpha(n_q, n_b)$ is bandwidth deficit for individual QoS flow. So, as n_q increases to n^*, $\alpha(n_q, n_b)$ goes to 1 and QoS flows have to borrow more time slots while best effort flows are starved. If n_b goes to ∞, $\alpha(n_q, n_b)$ goes to $\frac{\bar{b}}{h(n_q)}$ which is less than or equal to 1.

Since best effort flows lose their slots with probability $\alpha(n_q, n_b)$, the total capacity available to best effort flows is $g_b(n_q, n_b) = g(n_q + n_b)P_b(1 - \alpha(n_q, n_b))$, and so

$$g_b(n_q, n_b) = h(n_q + n_b)n_b \frac{h(n_q) - \bar{b}}{h(n_q) - h(n_q + n_b)}. \tag{7}$$

E. Capacity of best effort flows in normal regime

From (4), we can determine the lending probability $\beta(n_q, n_b)$:

$$\begin{aligned} \beta(n_q, n_b) \quad &= \quad \Big(h(n_q + n_b) - \bar{b}\Big) \frac{n_q + n_b}{\mathbb{E}[X^{(n_q)}|X^{(n_b)} \leq X^{(n_q)}]} \\ &= \quad 1 - \frac{\bar{b}}{h(n_q + n_b)}. \end{aligned}$$

Rearranging the above formula yields $(1 - \beta(n_q, n_b))h(n_q + n_b) = \bar{b}$, which is intuitively correct; each QoS flow balances its bandwidth to exactly $\bar{b}$.

Then, the capacity of best effort flows in the normal regime is

$$\begin{aligned} g_b(n_q, n_b) \quad &= \quad h(n_q + n_b)n_b + \\ &\quad \beta(n_q, n_b)\mathbb{E}[X^{(n_b)}|X^{(n_b)} \leq X^{(n_q)}]P_q. \end{aligned}$$

Using the same approach as in (5) and (6), we have

$$\begin{aligned} g_b(n_q, n_b) \quad &= \quad g(n_q + n_b)P_b + (1 - \frac{\bar{b}}{h(n_q + n_b)}) \times \\ &\quad \big(g(n_b) - P_b g(n_q + n_b)\big). \end{aligned}$$

Now, given the total capacity of best effort flows in overloaded and normal regime we can analyze 2-D Markov chain given in (1). To do so we only need to characterize $g(n)$ (or $h(n)$).

Example 1: In Rayleigh fading channel, signal strength has Rayleigh distribution, and random variable Y representing channel SNR has exponential distribution with mean ν^{-1}. Let

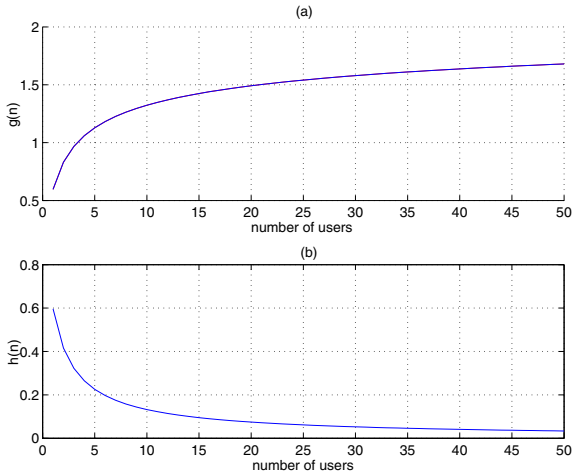

Fig. 1. Plot of the total capacity $g(n)$ (a) and individual capacity $h(n)$ (b) of 0dB Rayleigh fading channel.

$Z = Y^{(n)}$. Then, Shannon (or ergodic) capacity using max rate scheduling is calculated as

$$\begin{aligned} \mathbb{E}[X^{(n)}] \quad &= \quad \mathbb{E}[\log(1 + Z)] \\ &= \quad \int_0^\infty n(1 - e^{-\nu z})^{n-1} \nu e^{-\nu z} \log(1 + z) dz. \end{aligned} \tag{8}$$

Using the integral with incomplete gamma function $\Gamma(0, x)$,

$$\int_0^\infty e^{-kz} \log(1 + z) dz \quad = \quad \frac{e^k \Gamma(0, k)}{k} \tag{9}$$

where $\Gamma(0, x) = \int_x^\infty \frac{e^{-t}}{t} dt$ and using the binomial theorem, $g(n) = \mathbb{E}[X^{(n)}]$ is computed from (8) as

$$g(n) = n \sum_{k=0}^{n-1} (-1)^k \binom{n-1}{k} \frac{e^{\nu(k+1)} \Gamma\big(0, \nu(k+1)\big)}{k+1}.$$

Otherwise, $g(n)$ can be approximated by numerical calculation of (8). Fig. 1 shows an example of $g(n)$ and $h(n)$ in 0dB Rayleigh fading channel.

III. Stability

In this section we address stability of our system model. First we discuss the stability in case the maximum capacity of system is unbounded, *i.e.,* $g(n) \to \infty$ as $n \to \infty$. Ideal Rayleigh fading channel falls into this category. We show that the system is stable in this case. Then, we deal with the system that has finite capacity C and identify the necessary and sufficient condition of stability. Practical systems will of course fall in this second category.

A. Unbounded case

Theorem 1: If $g(n)$ is unbounded as n goes to ∞, *e.g.,* in ideal Rayleigh fading channel, then the system is stable for any offered load.

Proof: Let $\chi_t = \big(N_q(t), N_b(t)\big)$ denote the state of our irreducible, aperiodic, continuous time 2-D Markov chain on $\mathbb{S} = \{(n_q, n_b) | n_q \in \{0, \cdots, n^*\}, n_b \in \mathbb{Z}^+\}$. Based on Foster

theorem [23], since $g(n)$ is unbounded, for any offered load of best effort flows $\rho_b = \frac{\lambda_b}{\mu_b}$, we can find an $\exists l < \infty$ such that $g_b(n_q, n_b) > \rho$ for $\forall n_b > l$, which means the drift is negative with the corresponding Lyapunov function $\varphi(\chi_t) = n_b$. ■

B. Bounded case

Theorem 2: Suppose $C := \lim_{n\to\infty} g(n) < \infty$ and the maximum number of QoS users is limited to n^*. Then, the system is stable if and only if there exists $\exists l < \infty$ such that

$$\mathbb{E}[g(N_q(t), l)] > \rho_q(1-\eta_q)\bar{b} + \rho_b \tag{10}$$

where $\rho_q = \frac{\lambda_q}{\mu_q}$, $\rho_b = \frac{\lambda_b}{\mu_b}$ and $\eta_q = \pi_q(n^*)$, *i.e.,* blocking probability of QoS flows.

Proof: Since the number of QoS flows is bounded we need only consider the stability of best effort flows. We model this system as 1-D queueing system where the service rate is a state dependent random process $g_b(N_q(t), n_b)$.

The necessary condition is clear in that the total influx rate should be less than the average service rate. To prove the sufficient condition, we shall use the *Saturation Rule* [24]. Let T_{n_b} denote the time of the last departure from a system given it starts with n_b customers at time 0 and there are no more arrivals thereafter. The Saturation Rule says that T_{n_b} satisfies the strong law of large numbers, so $\lim_{n_b} \frac{n_b}{T_{n_b}}$ exists a.s. and the system is stable for the input process λ_b if

$$\lim_{n_b\to\infty} \frac{n_b}{T_{n_b}} > \lambda_b,$$

i.e., the departure rate for a saturated system exceeds the arrival rate. Let $T^l_{n_b}$ denote the stopping time from state $n_b > l$ to state $l < \infty$. Then, $T_{n_b} = T^l_{n_b} + T^0_l$. Since T^0_l is finite, $\lim_{n_b} \frac{n_b}{T_{n_b}} = \lim_{n_b} \frac{n_b}{T^l_{n_b}}$, a.s. Let s_i be the file size of best effort customer i. Since the total served bits on $[0, T^l_{n_b}]$ is less than $\sum_{i=1}^{n_b} s_i$,

$$\begin{aligned} \frac{n_b}{T^l_{n_b}} &\geq \frac{n_b \int_0^{T^l_{n_b}} g_b(N_q(t), l)dt}{T^l_{n_b} \sum_{i=1}^{n_b} s_i} \\ &= \frac{\frac{1}{T^l_{n_b}} \int_0^{T^l_{n_b}} g_b(N_q(t), l)dt}{\frac{1}{n_b}\sum_{i=1}^{n_b} s_i}. \end{aligned} \tag{11}$$

Then, $\lim_{n_b} \frac{1}{T^l_{n_b}} \int_0^{T^l_{n_b}} g_b(N_q(t), l)dt = \mathbb{E}[g_b(N_q(t), l)]$, a.s. since $g_b(N_q(t), l)$ is an ergodic process and $T^l_{n_b} \to \infty$, a.s. as $n_b \to \infty$ since service capacity is bounded by C. Also, by the strong law of large numbers, $\lim_{n_b} \frac{1}{n_b}\sum_{i=1}^{n_b} s_i = \mu_b^{-1}$, a.s. So, taking limit of (11) yields

$$\lim_{n_b} \frac{n_b}{T_{n_b}} = \lim_{n_b} \frac{n_b}{T^l_{n_b}} \geq \mathbb{E}[g_b(N_q(t), l)]\mu_b, \text{ a.s.}$$

Hence, if there exists $\exists l < \infty$ such that

$$\mathbb{E}[g_b(N_q(t), l)] > \rho_b, \tag{12}$$

then $\lim_{n_b} \frac{n_b}{T_{n_b}} > \lambda_b$. From (2), adding $\mathbb{E}[N_q]\bar{b} = \rho_q(1-\eta_q)\bar{b}$ to (12) makes (10) which completes the proof. ■

Corollary 1: Let $\bar{g}(n_q) := \lim_{n_b\to\infty} g(n_q, n_b)$. In the case where $g(n)$ is bounded, $\mathbb{E}[\bar{g}(N_q)]$ is less than C. So the maximum allowable influx rate into the system gets reduced by the integration of QoS and best effort flows.

Proof: Corollary 1 is obvious from the *capacity gap* described in the next section. ■

IV. Loss in opportunism

In this section we address the negative impacts of integrating QoS and best effort traffic. We shall refer to these as the *loss in opportunism* of service integration. The fundamental reason for losing throughput from opportunism comes from balancing QoS requirements vs. opportunism. To maximize system capacity we need to schedule users with high channel rate all the time. However, if we need to meet QoS requirements, at some time slots we are forced to select sub-optimal users resulting in loss of opportunism. Hence, we have a trade-off between guaranteeing QoS and maximizing capacity.

A. Capacity gap

Suppose that the system has a maximum capacity C and is supporting n_q QoS flows. Then, best effort traffic might expect its capacity to be $C - n_q\bar{b}$. However, the maximum opportunistic capacity that best effort flows achieve is $\bar{g}_b(n_q) := \lim_{n_b\to\infty} g_b(n_q, n_b)$. This limit is determined from (7) as $C\big(1 - \frac{\bar{b}}{h(n_q)}\big)$. So, we have a gap between $C - n_q\bar{b}$ and $\bar{g}_b(n_q)$. We call this quantity the *capacity gap*. Given n_q active QoS sessions each of which requires an average throughput of $\bar{b}$, the capacity gap $\xi(\bar{b}, n_q)$ is given by

$$\xi(\bar{b}, n_q) = \bar{b}n_q\Big(\frac{C}{g(n_q)} - 1\Big) > 0, \ n_q = 1, \ldots, n^*. \tag{13}$$

The importance of investigating this gap lies in that it affects not only the stability region of the system as stated in Corollary 1, but also degrades the performance of best effort flows because active QoS sessions deprive best effort sessions of available capacity, more than what is expected, resulting in local instability for best effort traffic. Local instability means that conditioned on a fixed number of QoS streams n_q, the arrival rate for best effort traffic λ_b exceeds the maximum service rate $\bar{g}_b(n_q)\mu_b$ and so traffic would temporarily accumulate. This usually happens when n_q remains high, and as a consequence best effort flows would experience long delays. Performance implications of this will be addressed in the next section.

Let us consider some characteristics of the capacity gap. From (13) we see that $\xi(\bar{b}, n_q)$ is proportional to the guaranteed bandwidth per QoS flow, and as a corner case, if $\bar{b} = g(1)$, then $n^* = 1$ and $\xi(1) = C - \bar{b}$, which means the system can support only one QoS flow and no best effort flows. The shape of $n_q\big(\frac{C}{g(n_q)} - 1\big)$ depends on $g(n)$. In turn, $g(n)$ is determined by the probability density function of the channel capacity. For example, for uniformly distributed channel capacity, $g(n) = C\frac{n}{n+1}$ and $\xi(\bar{b}, n_q) = \bar{b}$, which means the system experiences a constant capacity gap, *i.e.,*

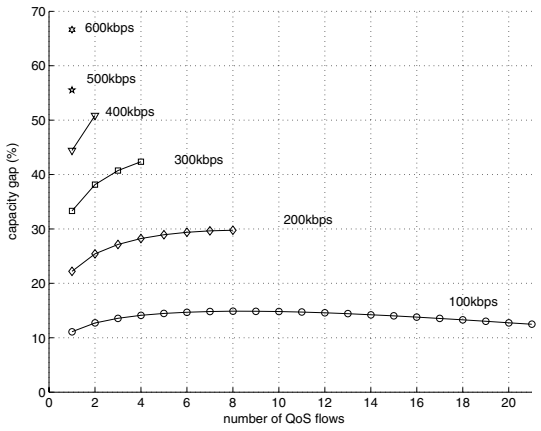

Fig. 2. Percentage of capacity gap $\frac{\xi(\bar{b},n_q)}{C}$ for the CDMA/HDR model describe in [1]: $\bar{b} = 100, \ldots, 600$Kbps.

independent of n_q. But usually the capacity gap will be an increasing function of n_q in the domain of interest. [3]

Example 2: This example will show that the capacity gap can be quite large in a real system. Let us define the *opportunistic gain* as the ratio of system capacity with and without opportunism, *i.e.,*

$$\kappa := \lim_{n\to\infty} \frac{g(n)}{g(1)} = \frac{C}{E[X]}. \tag{14}$$

Then, the capacity gap of introducing a single QoS user is given by

$$\xi(\bar{b}, 1) = \bar{b}(\kappa - 1).$$

Suppose that we have sufficient number of best effort users so that system capacity is close to C. In the CDMA/HDR system described in [1], $C = 2457$kbps and $E[X] = 659$kbps, so $\kappa = 3.75$. If we want to guarantee 300kbps per QoS flow, then capacity gap is $\xi(300, 1) = 300\times(3.75-1) = 825$kbps, which is 33 % drop in capacity from 2457kbps. For $\bar{b} = 200$kbps, the gap is 550kbps corresponding to 22% drop. Fig. 2 shows capacity gaps for various $\bar{b}$ and n_q. Note that each plot has different range of n_q for different $\bar{b}$ because the maximum number of QoS streams we can admit depends on $\bar{b}$. From the figure, we see that if $\bar{b} > 100$kbps, capacity gap increases as n_q grows but the slopes are decreasing in n_q. Thus, the capacity gap impacts the system mostly when the first few QoS flows are admitted. However, if $\bar{b}$ is small such as 100kbps, the capacity gap has a smooth peak and starts to decrease. This is because if $\bar{b}$ is small, the system can admit a sufficient number of QoS flows to generate the opportunistic capacity gain among the QoS flows. The next example illustrates the capacity gap of Rayleigh fading channels.

Example 3: Under a Rayleigh fading channel model, the impact of the capacity gap is more severe in low SNR than high SNR. Fig. 3 shows plots of capacity gap divided by $\bar{b}$ so as to only reflect the effect of n_q. We assume that a maximum of 1000 users can exist in the system and each SNR has a different bounded capacity. We see that the capacity gap of low SNR increases faster than that of high SNR. Considering the maximum capacity of low SNR is even less than high SNR, we see that the effect on capacity gap in the low SNR case is severe. This is consistent with what we see in the opportunistic gain from (14): κ is 3.57 at 0dB, 2.14 at 10dB, 1.62 at 20dB, 1.41 at 30dB. Thus we see that *high opportunistic gains will be associated with high loss in opportunism for integration.* This is further reflected in the delay performance considered in next subsection.

[3] If $\bar{b}$ is very small so n^* can be large enough, then it can be shown not to be an increasing function, *i.e.,* eventually decreases in n_q.

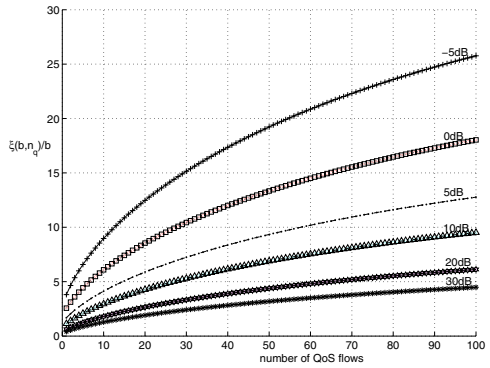

Fig. 3. Normalized capacity gap $\frac{\xi(\bar{b},n_q)}{\bar{b}}$ for various SNR under Rayleigh fading channel.

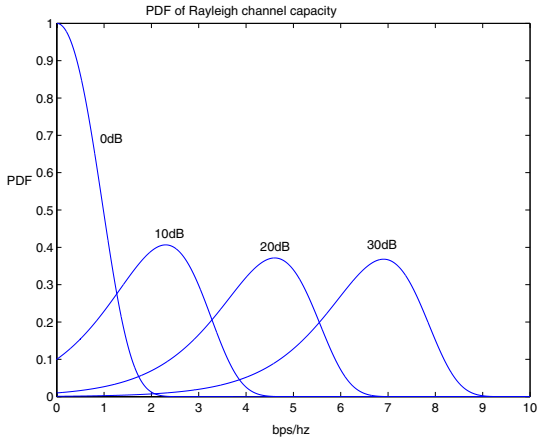

Fig. 4. Probability density function of Channel capacity under Rayleigh fading channel(bps/hz) for 0dB, 10dB, 20dB, 30dB.

B. Delay increase

In this section, we investigate the effect of the capacity gap on the delay performance of best effort flows. In a mixed user system, the ideal capacity that best effort flows could see under an opportunistic scheduling scheme would be

$$g_b^*(n_q, n_b) = g(n_q + n_b) - n_q\bar{b},$$

i.e., the overall maximum opportunistic capacity minus that given to QoS streams. However, the actual capacity seen in our

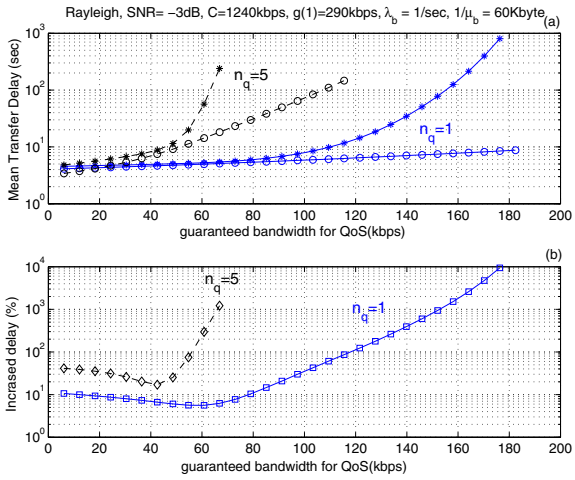

Fig. 5. Delay comparison at -3dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

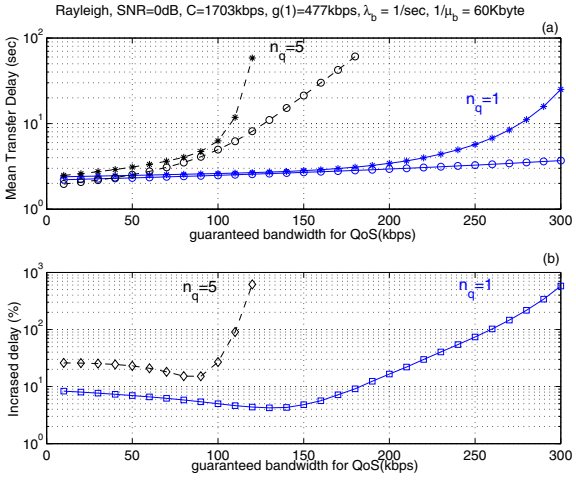

Fig. 6. Delay comparison at 0dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

model when attempting to meet QoS stream's requirements is $g_b(n_q, n_b)$, so the difference is $g_b^*(n_q, n_b) - g_b(n_q, n_b)$, which converges to the capacity gap as $n_b \to \infty$. In a dynamic system, as long as the best effort flows remain stable, offered load ρ_b will be served, yet delay will depend on the character of $g_b(n_q, n_b)$.

We will start, for simplicity, by evaluating the average delay of best effort users conditioned on a **fixed** number of active QoS sessions. In this case the distribution of best effort flows given n_q active QoS sessions is given by

$$\begin{aligned} \pi_b(n_b|n_q) &= \pi_b(0|n_q)\Pi_{i=1}^{n_b}\frac{\rho_b}{g_b(n_q,i)} \\ E[D|n_q] &= \frac{1}{\lambda_b}\sum_{n_b=1}^{\infty} n_b\pi_b(n_b|n_q). \end{aligned} \tag{15}$$

As a baseline delay performance, we substitute $g_b(n_q, n_b)$ with $g_b^*(n_q, n_b)$, and we will compare the delay under various SNR, $\bar{b}$ and n_q. We shall only consider the case where best effort flows are locally stable, *i.e.,*

$$\bar{g}_b(n_q) > \rho_b. \tag{16}$$

Even though the channel is Rayleigh fading, we assume that the range of X is finite since practical system can support only a finite number of users and they generate only finite opportunism. So, the delay of (15) is divided by $1 - \eta_b$ where η_b is blocking probability of best effort flows. We assume that up-to 1000 users can share the system.

Fig. 5 to Fig. 7 show the average delay for various $\bar{b}$ and $n_q = 1$ and 5. The figures show two curves, *'real'* and baseline delay associated with $g_b(n_q, n_b)$ and $g_b^*(n_q, n_b)$, respectively. Here, $\lambda_b = 1$/sec and $\mu_b^{-1} = 60$Kbytes as in [5]. The baseline delay represents an ideal delay performance under mixed traffic to show the delay penalty of the integration of QoS and best effort flows, conditioned on a fixed number of active QoS sessions.

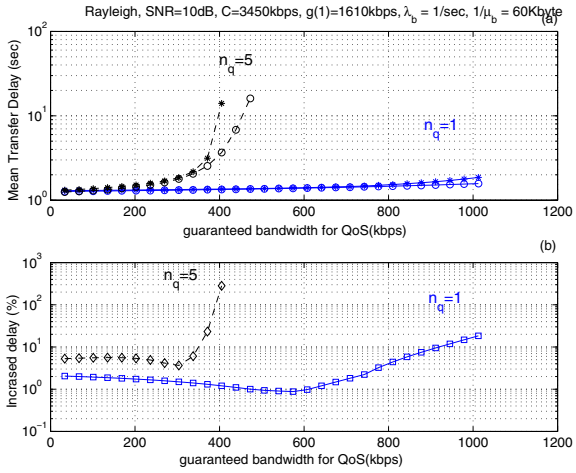

Fig. 7. Delay comparison at 10dB, Rayleigh fading channel. (a) real delay ($*$) and baseline delay ($\circ$). (b) Delay difference ratio.

In Fig. 5 (a) we see that as $\bar{b}$ grows, real and baseline delays increase and after some point the system is unstable. In Fig. 5 (b) we present delay difference ratio of real and baseline delays. Even for $n_q = 1$, we see penalty in performance. For example, with SNR $= -3$dB, $C = 1240$kbps, $g(1) = 290$kbps and $\bar{b} = 120$kbps, we see that real average delay is over 10 sec where as the baseline delay is around 5 sec. This is more than 100% increase. A single QoS flow can substantially increase average delays of best effort flows. If $n_q = 5$, the delay difference ratio grows more rapidly.

Comparing Fig. 5 through Fig. 7, one can see that the delay difference ratio is improved as the SNR increases. For example, if we compare SNR = -3dB, 0dB, 10dB at $n_q = 1$ and $\bar{b} = 0.4 \times g(1)$, then the ratios are 100% at -3dB, 15% at 0dB and 1% at 10dB.

Remark 4.1: From the above delay performance comparison, we can infer that guaranteeing bandwidth to a fixed

number of QoS users will make the delay of best effort longer than one might expect. This becomes severe for lower SNR, higher $\bar{b}$ and large n_q.

V. Admission control for best effort flows

In this section we consider admission control for best effort flows to improve delay performance. We propose a simple admission control strategy to reduce the impact of local instability where the number of best effort flows might temporarily grow.

A. Delay and local instability

In the previous section we considered the delay performance of best effort flows given a **fixed** number of active QoS sessions and saw that it deteriorates quickly in the number of QoS flows. Delays get higher as n_q increases and eventually may be locally unstable. Thus, if QoS flows remain in the system for a long time, best effort flows are not served much and their numbers may grow until best effort flows recover their capacity, *i.e.,* QoS sessions leave the system. As mentioned earlier this phenomenon is called *local instability* [13]. It is not unique to wireless network. It is common in bandwidth sharing systems where best effort flows are preempted by QoS flows. However, it is more serious in opportunistic wireless systems. Suppose that both of wired and wireless system have a maximum capacity C. For some given n_q, it is possible that a wired system does not experience local instability while wireless system does. This is because the wireless system needs a large number of flows to achieve opportunistic capacity. Clearly for every value of n_b, the capacity of best effort flows is smaller than that of a wired system. Furthermore, no matter how large n_b is, we have a capacity gap $\xi(\bar{b}, n_q)$ that prevents $g_b(n_q, n_b)$ from reaching the capacity of the wired network.

Fig. 8 illustrates a typical example of local instability for a Rayleigh fading channel at offered load / C = 0.79 with $\bar{b} =$ 100kbps, $C = 1.31$Mbps and $n^* = 10$. The joint distributions were computed numerically for the 2-dimensional Markov chain [25]. Since n_b is finite in a real system, this example assumes that n_b is limited by 30. Then, local instability results in accumulation of best effort flows at the boundary. In Fig. 8 (b) and (c) we see two peaks in the stationary distribution $\pi(n_q, n_b)$. The first peak at $(n_q, n_b) \simeq (2.5, 2.5)$ is preferred while the second peak at $(n_q, n_b) \simeq (7, 30)$ is problematic. For $\lambda_q = 0.023$/sec, $\mu_q^{-1} = 180$ sec, $\lambda_b = 1.3$/sec, $\mu_b^{-1} =$ 60Kbytes, we see that the state (n_q, n_b) is oscillating between two peaks along with drift arrow. If we decrease λ_b or λ_q, then the first peak becomes dominant and the second peak diminishes. Conversely, if we increase λ_b or λ_q, the first peak gradually disappears and the second peak dominates. This appears as to be a weak form of metastability, *i.e.,* where the system sees two operating regimes that are likely to jumps between them. This type of behavior is undesirable in practice, particularly if one of the modes corresponds to a poor performance, *e.g.,* long delay for best effort.

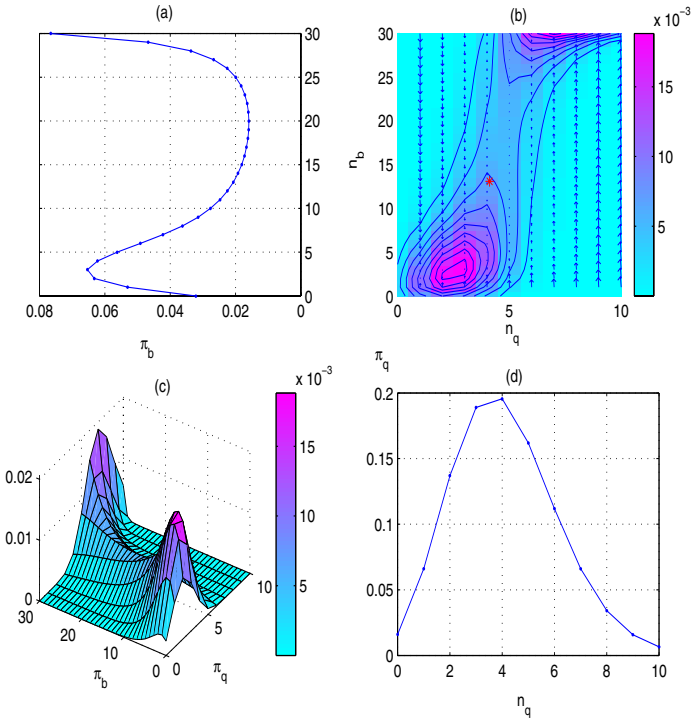

Fig. 8. An example of local instability (a) $\pi_b(n_b)$ (b) $\pi(n_q, n_b)$ contour and drift vector (c) $\pi(n_q, n_b)$ in 3-D view (d) $\pi_q(n_q)$: $\lambda_q = 0.023$/sec, $\mu_q^{-1} = 180$sec, $\lambda_b = 1.3$/sec, $\mu_b^{-1} = 60$Kbyte, load ratio = 0.79, $\bar{b} =$ 100kbps, $C = 1.31$Mbps, SNR 0dB Rayleigh fading channel.

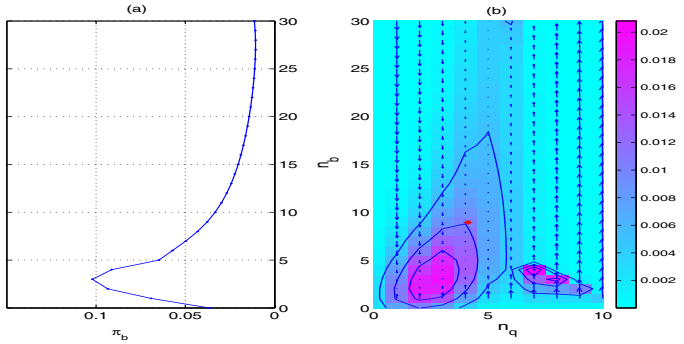

Fig. 9. An example of admission control applied for Fig. 8 for best effort flows (a) $\pi_b(n_b)$ (b) $\pi(n_q, n_b)$ contour and drift vector: $\theta = 0.5\mu_b^{-1}$.

B. Admission control for best effort flows

One way to preclude such *metastable* behavior is to ensure that the capacity of best effort flows is always greater than the offered load, *i.e.,* (16). However, this approach is not preferred because we need to block QoS flows before n_q reaches n^*. Another way is to apply admission control for best effort flows [14], [19]. For example, if the system state is (n_q, n_b), a new best effort flow might be blocked if

$$\rho_b > \theta + g_b(n_q, n_b + 1)$$

where θ is an admission control threshold. So, if the net influx rate of best effort flows is below some threshold, we admit all of them. As shown in Fig. 9 this *simple* admission control can

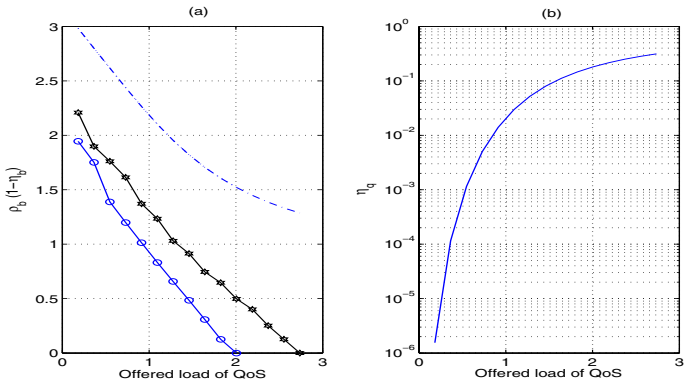

Fig. 10. Capacity region expansion by admission control at delay constraint = 20 sec for 10dB Rayleigh fading channel.(a) Admitted load of best effort flows with and without admission control: Ideal upper bound with no opportunism and no delay constraint (dashed), with CAC ($*$) and without CAC ($\circ$) (b) Call blocking probability of QoS flows

alleviate the performance impact of local instability effectively. We see that the undesirable peak is eliminated.

Another importance of admission control is the expansion of capacity region. We define the capacity region as the amount of admitted load of best effort flow under a given average delay constraint. Fig. 10 exhibits the capacity region and call blocking probability of QoS flows for a fixed offered load of QoS traffic $\rho_q \bar{b}$. The maximum average delay constraint is 20 sec. We see that the capacity region under the delay constraint is reduced relative to the theoretical limit of $C - \rho_q \bar{b}(1-\eta_q)$ assuming no capacity gap. Nevertheless the capacity region expands considerably. In this plot, the maximum number of best effort flows is assumed to be 100. The expansion of capacity region is meaningful, in that the service provider can in principle achieve more throughput under the same delay constraint and thus generate more revenue. Note that admission control strategy enhances the capacity region with an increased blocking probability for best effort flows.

VI. Conclusion

We have explored the flow-level dynamics and performance seen by a mixture of QoS and best effort flows sharing an opportunistic wireless system. In doing so, we proposed a new opportunistic scheduling scheme/model based on the concept of bandwidth borrowing/lending from/to best effort flows which enables the scheme to ensure a mean throughput in every time slot to QoS streams which is pertinent for the case where users see roughly homogenous channel variations. We evaluated the stability and the flow-level performance in this system. The results suggest that integrating QoS and best effort flows may degrade system performance in crucial aspects; reduction of the stability region, a gap in the capacity available to best effort traffic and increased file transfer delay. These negative impacts are referred to as *loss in opportunism*, and we found them to be proportional to the opportunistic gains, the guaranteed bandwidth, and to the number of QoS flows. We note, however, that these losses would be reduced in a system with a high SNR, assuming a Rayleigh fading channel model. Finally we explored the possibility of local instability, or a weak form of metastability in such systems, showing that it may indeed be more problematic in opportunistic wireless systems. To reduce this impact and ensure low delays for best effort flows, we suggest that admission control for best effort flows is perhaps even more critical than in wireline networks. We concluded the paper exhibiting the performance that a simple form of admission control would enhance.

Acknowledgment

The first author is grateful to Chan-Byoung Chae for helpful discussions.

References

[1] P. Bender and A. Viterbi, "CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Comm. Mag.*, pp. 70–77, July 2000.
[2] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficient-high data rate personal communication wireless system," *Proc. IEEE Veh. Technol. Conf.*, pp. 1854–1858, May 2000.
[3] Xin Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, pp. 451–473, 2003.
[4] Xin Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Jour. Select. Areas in Comm.*, vol. 19, pp. 2053–2063, Oct. 2001.
[5] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE Int. Conf. on Computer Comm. (INFOCOM)*, pp. 321–331, 2003.
[6] S. Patil and G. de Veciana, "Managing resource and quality of service in heterogeneous wireless systems exploiting opportunism," *IEEE/ACM Trans. Networking, accepted*, 2006.
[7] S. Patil and G. de Veciana, "Measurement-based opportunistic scheduling for heterogeneous wireless systems," *IEEE/ACM Trans. Networking, submitted*, 2006.
[8] S. Patil and G. de Veciana, "Feedback and opportunistic scheduling in wireless networks," *IEEE Trans. Wireless Comm., submitted*, 2006.
[9] D. Park, H. Kwon H. Seo, and B. G. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Trans. Comm.*, vol. 53, pp. 1919–1929, Nov. 2005.
[10] S. Shakkottai, "Effective capacity and QoS for wireless scheduling," *submitted for journal publication*, 2004.
[11] S. Shakkottai, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *American Mathematical Society Translations, Series 2, A volume in memory of F. Karpelevich, Yu. M. Suhov, Editor*, vol. 207, 2002.
[12] T. Bonald and A. Proutiere, "On performance bounds for the integration of elastic and adaptive streaming flows," *ACM SIGMETRICS*, pp. 235–245, 2004.
[13] F. Delcoigne, A. Proutiere, and G. Regnie, "Modeling integration of streaming and data traffic," *Performance Evaluation*, vol. 55, 2004.
[14] N. Benameur, S. Ben Fredj, S. Oueslait-Boulahia, and J. W. Roberts, "Quality of service and flow level admission control in the internet," *Computer Networks*, vol. 40, pp. 57–71, 2002.
[15] P. Key, L. Massoulie, A. Bain, and F. Kelly, "Fair internet traffic integration: network flow models and analysis," *Annals of Telecommunications*, vol. 59, 2004.
[16] E. Altman, A. Orda, and N. Shimkin, "Bandwidth allocation for guaranteed versus best effort service categories," *Queueing Systems*, vol. 36, pp. 89–105, 2000.
[17] T-J. Lee and G. de Veciana, "Model and performance evaluation for multiservice network link supporting ABR and CBR services," *IEEE Commun. Letters*, vol. 4, 2000.
[18] L. Massoulie and J.W. Roberts, "Bandwidth sharing and admission control for elastic traffic," *Telecommunication Systems*, vol. 15, pp. 185–201, 2000.
[19] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslait-Boulahia, and J.W. Roberts, "Integrated admission control for streaming and elastic traffic," *QofIS*, pp. 69–81, 2001.
[20] R. Knopp and P.A. Humblet, "Information capacity and power control in single-cell multiuser communications," *Proc. IEEE Int. Conf. on Comm.*, pp. 331–335, 1995.
[21] B. S. Tsybakov, "File transmission over wireless fast fading downlink," *IEEE Trans. Info. Theory*, vol. 48, pp. 2323–2337, Aug. 2002.
[22] D. Bertsekas and R. Gallager, "Data networks," *Prentice Hall*, 1992.
[23] G. Fayolle, V. A. Malyshev, and M. V. Menshikov, "Topics in the constructive theory of countable markov chains," *Cambridge University Press*, 1985.
[24] F. Baccelli and P. Bremaud, "Elements of queueing theory," *Springer-Verlag*, 2004.
[25] M. F. Neuts, "Matrix-geometric solutions in stochastic models," *The Johns Hopkins University Press*, 1981.