Opportunistic Scheduling for Streaming Multimedia Users in High-Speed Downlink Packet Access (HSDPA)

Arsalan Farrokh, *Member, IEEE*, and Vikram Krishnamurthy, *Fellow, IEEE*

***Abstract*—High-speed downlink packet access (HSDPA) achieves high data rates and high spectral efficiency by using adaptive modulation and coding schemes and employing multicode CDMA. In this paper, we present opportunistic algorithms for scheduling HSDPA users and selecting modulation/coding and multicode schemes that exploit channel and buffer variations to increase the probability of uninterrupted media play-out. First, we introduce a stochastic discrete event model for a HSDPA system. By employing the discrete event model, we transform the scheduling problem of providing uninterrupted play-out to a *feasibility* problem that considers two sets of stochastic Quality-of-Service (QoS) constraints: *stability* constraints and *robustness* constraints. A methodology for obtaining a feasible solution is then proposed by starting with a so-called *stable* algorithm that satisfies the stability QoS constraints. Next, we present Stochastic Approximation algorithms that adapt the parameters of the stable algorithm in a way that a feasible point for the robustness QoS is reached within the feasibility region of the stability QoS.**

***Index Terms*—Adaptive modulation and coding (AMC), Hybrid ARQ (H-ARQ), high-speed downlink packet access (HSDPA), Markov Fading Channel, opportunistic scheduling, Quality-of-Service (QoS), stability, stochastic approximation, UMTS.**

I. Introduction

THE IMT-2000 standards for third-generation (3G) wireless networks, released in 1999, promise high bandwidth efficiency for supporting a mix of real-time and high data rate traffic. Three of the five standards proposed in IMT-2000, including UMTS (the European contribution), are based on wideband code division multiple access (WCDMA). Though, WCDMA/UMTS specifications fully meet the IMT-2000 requirements for 3G networks, there is still an increasing demand for much higher downlink data rates along with a better Quality-of-Service (QoS). In order to meet these demands, a new packet concept, called high-speed downlink packet access (HSDPA), was recently published in Release 5 of 3GPP UTRA-FDD specifications [1]. HSDPA is an extension of the UMTS standard that provides an improved peak data rate and an increased throughput for world-wide cellular systems. The main features that collectively describe HSDPA are: adaptive modulation and coding (AMC) schemes, Turbo codes, higher order modulation (16-QAM), CDMA multicode operation, fast physical layer hybrid automatic repeat request (ARQ), and short transmission time intervals. These features enable HSDPA to support transmission rates of up to 10.7 Mbit/s [2] and allow high-rate downlink transmission of streaming media and data for large numbers of users in a cellular system.

HSDPA is a time-slotted CDMA system and requires a suitable scheduling policy to achieve the desired performance. A transmission scheduler, based on the channel information of each user, must decide which user should be scheduled at each time slot and furthermore, what type of modulation, coding, and multicode (MCM) combination should be used. Scheduling methodologies are not specifically defined as part of the HSDPA standard. This motivates the need to develop efficient scheduling algorithms that exploit the nature of wireless channel to achieve the required QoS for the users.

Literature Review: 3G wireless traffic (e.g., real-time multimedia) can be very resource demanding and hence employing suitable radio resource allocation (RRA) schemes is an important aspect of a 3G wireless system design. In [3] and [2], link and network layer performance aspects of 3G WCDMA/HSDPA systems in terms of improving peak data rates and blocking probabilities are studied. The retransmission scheduling is also considered and simulated at both link and network layers. The study shows a tradeoff between user fairness and the cell throughput. In [4], and [5], opportunistic scheduling methods are used for a time-slotted wireless system to optimize the cell throughput while maintaining certain fairness among the users. Two categories of fairness in terms of minimum-throughput guarantee and specific shares of scheduling times are considered. However, in these papers there is no assumption on the status of the user buffers. Assuming a linear relationship between the bit rate and signal-to-noise ratio, it is shown in [6] and [7] that scheduling one user at a time with full power gives a better throughput performance compared to scheduling simultaneous users. In [8], the problem of scheduling CDMA users (one user at a time) in downlink with variable channel conditions is considered. The authors have proposed an algorithm that considers both channel state information and user buffer state information for scheduling decisions. The algorithm, in general, guarantees the QoS but may not achieve the system capacity. A general approach to support QoS of multiple real-time data users is proposed in [9]. The proposed scheduling algorithm of this paper is similar to [8] and can be used to maximize the number of users that can be supported with the desired QoS.

Manuscript received December 16, 2004; revised November 4, 2005. This work was supported by a NSERC strategic research grant. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wanjiun Liao.

The authors are with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: arsalanf@ece.ubc.ca; vikramk@ece.ubc.ca).

Digital Object Identifier 10.1109/TMM.2006.876227

Main Results: In this paper, we consider scheduling algorithms for streaming multimedia users in a HSDPA transmission system. The main results of this paper are as follows.

i) In Section II, we introduce a discrete-event stochastic model for streaming users in HSDPA. The key features of HSDPA system as described in 3GPP standards are included in our proposed model. Furthermore, a fading channel, modeled by a finite state Markov process, is considered.

ii) In Section III, we formulate the scheduling problem of providing uninterrupted play-out as a stochastic *feasibility* problem. First, QoS constraints are formulated in the form of *long term* throughput constraints and stochastic *robustness* constraints of the mobile-station buffers. These constraints collectively define the feasibility region of the scheduling problem. Next, by exploiting the structure of the model proposed in (i), equivalent QoS constraints for the base-station buffers are formulated. Long term throughput performance is formulated as the *stability* constraints on the base-station buffers (queues). Roughly speaking, by the stability of a queue we mean the condition that the queue content does not grow to infinity (the precise meaning of the stochastic stability of a queue will be explained in Section III). The robustness QoS constraints are formulated as *probabilistic* constraints on the base-station buffers (queues). Note that the terms "buffer" and "queue" are used interchangeably in this paper, however the former refers to a physical entity while the latter describes a mathematical model. Furthermore, by the term "buffer size", we mean the size of the physical buffer (maximum length of the queue) and by the term "queue length" we mean number of the existing customers (bits) in the queue.

iii) In Section IV, we present a joint opportunistic user-scheduling and MCM assignment algorithm as a solution to the HSDPA scheduling problem. Section IV-A contains a complete summary of the proposed opportunistic scheduling algorithm. We first propose a framework for assigning MCM in HSDPA in a way to maximize the achievable bit rates of the scheduled users. We then present a user-scheduling method by *customizing* a particular scheduling algorithm denoted by modified largest-weighted-(unfinished) work-first (M-LWWF) [8]. The M-LWWF algorithm schedules only one user at a time and provides a simple algorithm with relatively low complexity. It is analytically proven that the proposed customized algorithm provides us with a feasible long term solution (if such a long term feasible solution exists). However, additional degrees of freedom are embedded in the M-LWWF algorithm by assigning *fairness* parameters to the users. The scheduling priority across the users can be further controlled by adjusting the fairness parameters to satisfy additional constraints. We propose a stochastic-approximation-based adaptive algorithm that dynamically (i.e., in real time) adjusts each user's weight in a way that both robustness and stability QoS constraints are satisfied.

iv) In Section V, we simulate the HSDPA transmission for streaming users across a finite state Markovian fading channel. In our simulations the relation between signal-to-noise ratio (SNR) and frame error rate (FER) is extracted from the data provided by Motorola research [10]. The data in [10] accounts for the extensive use of Turbo codes in HSDPA systems.

II. HSDPA Streaming Discrete Event System Model

In this section, a stochastic discrete event system model for streaming users in HSDPA is presented. Here, the term BS refers to the *base-station* and the term UE refers to the *user-equipment* (i.e., mobile station). Throughout this paper we use superscript index i to denote the i'th component of a vector and subscript index $k \in \mathcal{Z}_+$ to denote the discrete time slot index, where $\mathcal{Z}_+ = \{0, 1, 2, \ldots\}$ is the set of non-negative integers.

It is convenient to outline our HSDPA streaming model in terms of the following nine elements [2], [3]. We consider L users and define $\mathcal{I}$ to be the set of all users: $\mathcal{I} := \{1, 2, \ldots, L\}$.

1) Transmission Time Interval (TTI): Time is the resource that is shared among the HSDPA users. The time axis is divided into slots of equal duration referred to as TTI. Define

$$\Delta T = \text{Duration of one TTI.} \tag{1}$$

By convention, time slot (or discrete time) $k, k \in \mathcal{Z}_+$ is the time interval $[k\Delta T, (k+1)\Delta T)$. As explained in Section II-A-2 of introduction, *we assume only one user is scheduled at each time slot and scheduling decisions are made at times* $k\Delta T$, $k \in \mathcal{Z}_+$.

2) Power: Fast power control feature of the WCDMA system is disabled in the HSDPA standards (3GPP UTRA-FDD). Therefore, we assume that transmission power is fixed for all time slots.

3) Fading Channel: Channel quality for user i at time k is characterized by the average received SNR of the user i at the k'th time slot. Note that since the transmission power is fixed, SNR can be used as a direct measure for the channel quality. The channel quality indicator (CQI) or simply the *channel state* is defined as the collection of the channel qualities for the L users at time k and is denoted by vector c_k

$$c_k = \left(c_k^1\ c_k^2\ \ldots\ c_k^L\right) \tag{2}$$

where c_k^i is the channel quality of user i at time k. Assume $c_k \in S$ is an irreducible discrete time Markov chain with state space S. State space $S := \{s_1, s_2, \ldots, s_M\}$ is obtained by quantizing the range of the received instantaneous SNRs into M vector levels. Here, s_i are row vectors with L elements that define the quantization levels. We assume that CQI vector is measured by the users and reported to the the BS at each time slot. Therefore, the BS has a perfect knowledge of vector c_k for $k \in \mathcal{Z}_+$.

Remark on the Markovian Channel Assumption: In general, any stationary continuous time process can be quantized into a finite state irreducible Markov process. In particular, there are well-known methods for modeling individual Rayleigh fading channels by irreducible finite-state Markov processes (FSMP) [11], [12]. A finite-state model of channel is needed to formalize the proof of the queues stability. As can be seen in Algorithm 1,

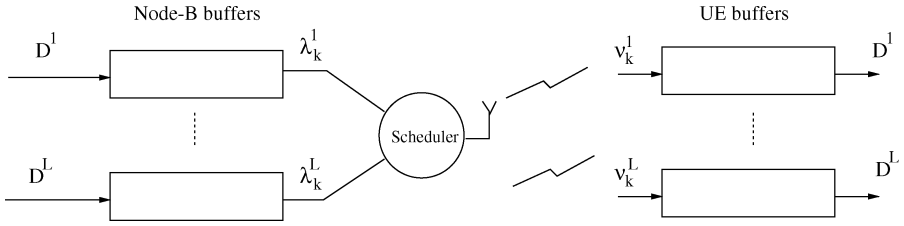


Fig. 1. Wireless streaming media system model.

in Section IV-A, the scheduling decision at any given time is made based on the channel state at that time slot. However, since at any given time slot, the channel state is fully observed, the number of iterations per step or the complexity of the algorithm does not depend on the number of the channel states. In this view, any number of channel states (M) according to the desired "resolution" may be selected without any penalty in computational complexity.

4) Adaptive Modulation and Coding (AMC): In HSDPA, different modulation and error-correcting coding techniques can be dynamically selected from a set of modulation and coding schemes (MCS). The reasoning for using the adaptive MCS can be explained by examining two extreme situations for the channel. First, consider a situation that channel experiences a relatively low quality (low SNR and high interference level). In this case, instead of increasing power (note that there is no power control in HSDPA), the transmission scheduler selects a more "robust" modulation scheme (such as QPSK) that will operate well in lower E_b/N_0 (bit energy to noise density). Furthermore, error coding techniques such as Turbo code, with more redundancy will be employed to handle the lower SNR. On the other hand, if the channel enjoys a relatively high SNR, then the scheduler selects a more "aggressive" (in terms of the data rate) modulation scheme such as 16QAM. We assume each MCS is represented by a MCS number such that

$$\text{MCS number} \in \mathcal{M} := \{1, 2, \ldots, |\mathcal{M}|\}. \tag{3}$$

In the case of scheduling user i at time k

$$m_k^i := \text{MCS number of user } i \text{ at time k} \tag{4}$$

where $|\mathcal{M}|$ denotes the cardinality (number of elements) of set $\mathcal{M}$ and $m_k^i \in \mathcal{M}$.

5) Multicode Operation: In HSDPA, multicode CDMA is used at each time-slot for transmission to the scheduled user. By choosing a larger set of spreading codes, higher data rates can be achieved, however, higher SNR will be required for transmission with more codes. Depending on the UE capabilities, different number of simultaneous spreading codes can be selected (e.g., 5, 10, 15). Define $\mathcal{N}$ to be the set of supported spreading codes for user i. In the case of scheduling user i at time k

$$n_k^i := \text{Number of spreading codes of user } i \text{ at time } k. \tag{5}$$

where $n_k^i \in \mathcal{N}$.

6) Transmission Rate Set: We refer to the rate chosen by the BS to transmit the data to the UE as the *transmission rate*. The transmission rate at each time slot k is fixed and chosen from a finite set of possible transmission rates by the BS. Define

$$r_k^i := \text{The (achievable) transmission rate to user } i \text{ at time } k. \tag{6}$$

Let R_T be the finite set of all possible transmission rates from the BS to the UE so that $r_k^i \in R_T$, $i \in \mathcal{I}$, $k \in \mathcal{Z}_+$. We assume that a unique transmission rate is associated with any combination of MCS and multicode. Let $r : \mathcal{M} \times \mathcal{N} \to R_T$ be a function that maps the MCS and multicode combination to the corresponding transmission rate, so that

$$r\left(m_k^i, n_k^i\right) := r_k^i, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+. \tag{7}$$

7) Physical-Layer Hybrid ARQ (H-ARQ): ARQ is one of the error-control procedures used in data communication systems. If the user receiver detects transmission errors in a frame, it will automatically request retransmission from the BS. The BS retransmits the frame until it receives an acknowledgement (ACK) from the receiver or the error persists beyond a predetermined number of retransmissions. By a *frame* of data we mean the data that is delivered at any time slot (TTI). Note that the length of a frame in terms of bits is varying due to the different data rates at different time slots. H-ARQ uses a code with more redundancy that will correct some frame errors and hence reduces the average number of retransmissions. H-ARQ operation in HSDPA is relatively fast since it is performed in the physical layer between the base station and the user equipment.

We define f^i as a function that gives the *instantaneous* frame error rate (FER) of user i in terms of MCS, number of codes, and channel state of user i. The term error rate here is used to denote the error probability. By the instantaneous FER, we mean the *true* FER that indicates the probability of the occurrence of a frame error at the given time slot. Define

$$\begin{aligned} f_k^i &= \text{instantaneous FER at time } k \text{ for the} \\ &\quad \text{scheduled user } i \\ &= f^i\left(m_k^i, n_k^i, c_k^i\right). \end{aligned} \tag{8}$$

8) Buffers and Streaming Rates: Fig. 1 shows a queuing model for streaming HSDPA users. The streaming system is modeled with symmetric buffers, i.e., each UE buffer has an equivalent buffer in the BS with the same size. Assume that the data in the UE buffer is discharged with the constant rate D^i

which is referred to as the user i *play-out* rate. Furthermore, assume the incoming data rates from the server are also constant. Let $\bar{D}^i$ denote the arriving data rate to the user i buffer in the BS (incoming rate from the server): clearly, since $\bar{D}_i$ and D_i are both constants, we need to assume $\bar{D}_i = D_i$ in a stable system model. Note that the above model can be extended to the case where the arrival rate from the server is not constant. In general, the results of this paper, i.e., stability and robustness of the BS buffers under the proposed opportunistic algorithm, are valid if the arrival rate from the server is any *Renewal* process [8]. However, this assumption poses more formal technicalities and renders the *exact* analysis of our proposed model intractable.

Define the user i BS queue length as

$$V_k^i := \text{Number of bits in the user } i \text{ BS buffer at time } k \quad (9)$$

Also, define the collection of buffer states for L users as

$$V_k = \left\{V_k^1, V_k^2, \ldots, V_k^L\right\}. \quad (10)$$

Similarly, the number of bits in the buffer of UE i at time k is denoted by U_k^i. Let $\bar{B}^i$ be the full size of the BS (or the UE) buffer of user i. The relation between V_k^i and U_k^i is then given by

$$V_k^i + U_k^i = \bar{B}^i, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \quad (11)$$

where we have assumed $V_0^i = 0$ and $U_0^i = \bar{B}^i$. Let ν_k^i be the arriving data rate (bit/s) to the buffer of UE i at time k. Also, let λ_k^i be the rate that the data is discharged from the buffer i of the BS at time k. By convention, we assume if a frame error occurs in time slot $k = s$ then $\lambda_s^i = 0$. This is reasonable since the BS does not transmit a new frame to the UE till it receives an ACK confirming a successful previous transmission. The effective transmission rate at time s is then zero because the same packet will be retransmitted at the next scheduling time. We conclude that the two rates can be regarded to be identical: $\nu_k^i = \lambda_k^i$. The evolution of the BS and the UE buffer content for user i is then given by the two following equations:

$$V_{k+1}^i = V_k^i - \lambda_k^i \Delta T + D^i \Delta T, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \quad (12)$$

$$U_{k+1}^i = U_k^i + \lambda_k^i \Delta T - D^i \Delta T, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \quad (13)$$

where ΔT is defined in (1).

9) Per-TTI (Instantaneous) Capacities: Recall that HSDPA system performs ARQ in the *physical layer*. This implies that if a frame error occurs at any given time slot, the entire frame will be retransmitted in the next scheduled time slot. In order to quantify the achievable throughput to user i at time k, we consider a notion of *per-TTI (instantaneous) capacity* that incorporates the event that a frame error occurs. Define a random variable μ_k^i as

$$\mu_k^i := \text{User } i \text{ per} - \text{TTI capacity at time } k. \quad (14)$$

This means, in the event of scheduling user i at time k, user i receives $\mu_k^i \Delta T$ bits at time slot k (or the entire queue i content whichever is smaller). In terms of transmission rate, we have

$$\mu_k^i = r_k^i \mathbf{1}_{\{\text{No Frame Error at time } k\}}, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \quad (15)$$

where r_k^i is defined in (6) and $\mathbf{1}_{\{A\}}$ is the indicator function of the event A (it is equal to 1 if A occurs, otherwise is zero). As can be seen in (15), if a frame error occurs in a time slot, the per-TTI capacity is defined to be zero at that time slot.

III. Formulation of HSDPA Scheduling Problem With QoS Requirements

In this section, we consider the problem of opportunistic joint user scheduling and MCM assignment in HSDPA. The objective is to enable a smooth play-out by applying a suitable user-scheduling and MCM assignment policy.

First, the general structure of stationary opportunistic policies is defined. The term *opportunistic* refers to the fact that the channel and/or buffer (state) variations are exploited by the scheduler to assign *relatively better* users (in a sense to be defined below).

Next, using the discrete event system model described in Section II, the condition of smooth play-out is formulated as a *feasibility* problem with stochastic robustness and stability QoS constraints. Informally speaking, the objective is to keep the UE buffers *non-empty* in a long run and limit the variations of the buffers to avoid any interruption in the media play-out.

Main Result: The main result of this section is the formulated problem in (25)–(27) which states the required QoS constraints for streaming users in HSDPA.

A. Opportunistic Joint User Scheduling and MCM Assignment

In this subsection, we define a general structure for opportunistic joint user scheduling and MCM assignment policies in HSDPA. The user scheduling and MCM assignment are dynamically performed based on the system state information at any time slot. The *state* of the system can be defined as the collection of the channel and the buffer states

$$x_k := (c_k, V_k) \in \mathcal{X} \quad (16)$$

where c_k and V_k are defined in (2) and (10), respectively, and $\mathcal{X}$ is the state space. For now assume x_k is a stationary and ergodic Markov process. Later, based on the proposed scheduling algorithm and assuming stability of the queues, this assumption is justified.

A *scheduling policy* ϕ is defined as a function that maps the system state to a system user at a generic time slot. The mapping $\phi(x) = j, \phi : \mathcal{X} \to \mathcal{I}$ means that user j is scheduled if the state is reported to be x. As seen above, we assume only one user is scheduled at any given time slot. Furthermore, we only consider the class of stationary policies and hence ϕ is not a function of time. These assumptions will be justified later in Section III-C, where we formulate the scheduling problem as a feasibility problem.

An MCM assignment rule ψ is defined as a function that maps the system state to a pair of multicode and MCS number at any time slot

$$\psi(x) = (m, n), \quad \phi : \mathcal{X} \to \mathcal{M} \times \mathcal{N} \tag{17}$$

where (17) means that MCS number m along with n multicode is chosen for the scheduled user if the state is reported to be x.

A joint user scheduling and MCM assignment policy Λ is defined as the combination of the scheduling policy ϕ and MCM assignment rule ψ, in a generic time slot:

$$\Lambda = \{\phi, \psi\} \tag{18}$$

Hence, if user $j \in \mathcal{I}$ is scheduled at time k, we have: $\Lambda(x_k) = (j, m_k^j, n_k^j)$ where m_k^i and n_k^i are defined in (4) and (5), respectively, for all $i \in \mathcal{I}$.

B. Formulation of QoS Constraints for HSDPA Streaming

Here, we state the QoS constraints that enable the smooth media play-out for streaming HSDPA users.

Stability QoS: To have a smooth (uninterrupted) play-out, the user buffer content (in the UE) must have a non-zero steady-state regime. In other words, in a long run the user buffers in the UE must remain *non-empty*. Our streaming model, based on (11), (12), and (13), allows us to express an equivalent requirement for the users buffers in the BS. (11) asserts that an empty buffer in the UE corresponds to a full buffer in the BS. Hence the situation that a UE buffer i is "under-running" corresponds to the situation that the BS buffer i is "blowing up". It can be concluded that for having uninterrupted play-out in a long run, the users queues in the BS must be *stable*. By stability we mean the existence of a stationary regime for the queue content process so the content of the queues does not show the tendency to grow to infinity [8]. In other words, the steady-state probability distribution of the queue length process V_k^i should be well defined. This condition is expressed as follows (with $\mathbf{P}$ denoting the probability measure):

$$\lim_{k\to\infty} \mathbf{P}\left(V_k^i \le \theta\right) = F_i(\theta), \quad i \in \mathcal{I} \tag{19}$$

where $F_i(\cdot)$ is a valid distribution function (i.e., $F_i(\cdot) \ge 0$ is right continuous and $\lim_{\theta\to\infty} F_i(\theta) = 1$), and V_k^i is defined in (9).

Remark: In the following, we assume only the steady-state behavior of the queues, i.e., time $k \in \mathcal{Z}_+$ is sufficiently large.

By formulating equivalent QoS constraints for user buffers in the BS (rather than for the UE buffers) we will be able to use certain stochastic stability results from [8]. First recall that the net discharging rate of queue i (in the BS) at time k is denoted by λ_k^i. Therefore λ_k^i is equal to the user i throughput at time k and is given by

$$\lambda_k^i = \min\left\{\mu_k^i \mathbf{1}_{\{\phi(x_k)=i\}}, \frac{V_k^i}{\Delta T}\right\}, \quad i \in \mathcal{I} \tag{20}$$

where μ_k^i is given by (15) and ϕ is the scheduling policy. The equality in (20) holds because in the event that the available rate completely discharges the BS buffer during one time slot, the effective data rate of user i during that time slot is equal to $V_k^i/\Delta T$. The general expression for λ_k^i is then described as in (20). For stable queues the arriving rate D^i matches the net discharging rate λ_k^i in a long run. Assuming a scheduling policy ϕ under which the system is stable. i.e., the discrete Markov chain x_k (state) is ergodic, we have (with $\mathbf{E}$ denoting the expectation operator)

$$\mathbf{E}\left(\lambda_k^i\right) = D^i, \quad i \in \mathcal{I}. \tag{21}$$

The sufficient and necessary condition for (21), i.e., stability of user queues, is given by (see [8, Th. 1])

$$\mathbf{E}\left(\mu_k^i \mathbf{1}_{\{\phi(x_k)=i\}}\right) > D_i, \quad i \in \mathcal{I}. \tag{22}$$

Robustness QoS: To allow an uninterrupted real-time streaming of the users, the situation of an empty or under-run buffer must be avoided. Condition (22) ensures the existence of a long-term (stationary) non-zero process for the UE buffer but it does not address the variations of the UE buffer content. It is possible that under a stable scheduling policy individual users often encounter an empty or under-run buffers. To avoid this situation, a stochastic robustness QoS is considered as follows: the probability of having an under-run buffer at each time slot needs to be smaller than a certain threshold. This constraint can be formulated more generally as

$$\mathbf{P}\left(U_k^i < \theta_i\right) \le \delta_i, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \tag{23}$$

where θ_i is a threshold level for the UE buffer of user i and δ_i is the threshold probability for user i. Using (11), an equivalent constraint can be written for the BS buffers

$$\mathbf{P}\left(V_k^i > \eta_i\right) \le \delta_i, \quad i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \tag{24}$$

where $\eta_i = \bar{B}_i - \theta_i$ is a threshold level for the BS buffer of user i and δ_i is the threshold probability. Recall that $\bar{B}_i$ is the full BS buffer size of user i given in (11) (which is also equal to the full UE buffer size of user i by the symmetric-buffers model shown in Fig. 1).

C. HSDPA Scheduling Problem

In this subsection, based on the QoS constraints introduced in Section III-B, we formulate the scheduling problem for streaming users in HSDPA as a *feasibility* problem. A more general problem can be stated as an *optimization* problem (i.e., minimizing a certain cost function) with the stability and robustness QoS constraints. This problem can be generally formulated as a special MDP denoted by average-cost-per-stage infinite horizon problems [13]. The solution to this MDP can be obtained by Dynamic Programming techniques and may require scheduling more than one user at a given time slot.

However, as the number of buffer and channel states increases, the complexity of the MDP grows exponentially and renders the solution impractical [1]. In this paper by trading-off optimality with feasibility, we will be able to obtain a practical opportunistic algorithm with low complexity that assumes only one user is scheduled at any time slot. Also note that similar to infinite horizon MDPs, the scheduling policy is assumed to be stationary otherwise no simple formulation of the problem exists [13]. The feasibility problem can be stated as the collection of the robustness and stability constraints introduced in (24) and (22), respectively. An additional constraint is also considered in (27) which asserts that the instantaneous FER for each user needs to be kept below a specific level f^i_{max} at each time slot (if the channel quality allows). The constraint on the maximum instantaneous FER prevents excessive retransmissions which may cause certain technical issues. The scheduling problem for streaming users can be formulated as follows.

Problem: Find a joint scheduling and MCM assignment policy $\Lambda = \{\phi, \psi\}$ [defined in (18)) to satisfy the following QoS constraints for HSDPA users]:

$$\mathbf{E}\left(\mu^i_k \mathbf{1}_{\{\phi(x_k)=i\}}\right) > D_i \quad \text{(Stability constraints)} \tag{25}$$

$$\mathbf{P}\left(V^i_k > \eta_i\right) \le \delta_i \quad \text{(Robustness constraints)} \tag{26}$$

$$f^i_k \le f^i_{max}$$

$$i \in \mathcal{I}, \quad k \in \mathcal{Z}_+ \tag{27}$$

where μ^i_k is given by (15) and f^i_k is defined in (8).

IV. Opportunistic Scheduling Algorithm for Streaming HSDPA Users

In this section we present an opportunistic algorithm for joint user scheduling and MCM assignment as a solution to the problem outlined in (25)–(27). First, we outline the methodology to obtain a feasible solution and summarize the steps of the algorithm. Next, each step of the algorithm is discussed in more detail.

Main Result: The main result of this section is *Algorithm 1* in Section IV-A, which is proposed as a solution to the HSDPA scheduling problem. *Remark*: If not specified, throughout this section, we assume $i \in \mathcal{I}$, $k \in \mathcal{Z}_+$.

A. Summary of the Opportunistic Streaming Algorithm

We consider the following *general structure* for the HSDPA scheduling algorithm:

$$\phi(x_k) = \arg\max_{i \in \mathcal{I}} \gamma_i V^i_k \mu^i_k \tag{28}$$

where μ^i_k are the per-TTI capacities defined in (14) and V^i_k are the queues' lengths defined in (9). γ_i are real constants, denoted by *fairness parameters*, that provide additional flexibility for assigning priorities (fairness) among the users. The idea is to assign higher priorities to the relatively better channels (higher μ^k_i) and hence exploiting channel variations to maximize the overall throughput. On the other hand, the algorithm assigns higher priorities to the users with larger queue lengths and therefore reducing the probability of the queues growing larger. The fairness parameters can then be adjusted to further shape the priority distribution among the users. The algorithm in (28) is denoted by modified largest-weighted-(unfinished) work-first (M-LWWF) and is proven to guarantee the stability of the queues (if possible with any other policy) [8]. We present a joint opportunistic scheduling and MCM assignment algorithm by further modifying the M-LWWF scheduling algorithm and combining it with a suitable MCM assignment scheme (in particular, the per-TTI capacities in M-LWWF are replaced with their optimal minimum-variance estimates). We prove that the resulting joint opportunistic scheduling and MCM assignment algorithm guarantees the stability of the queues as well and hence satisfies the stability condition in (25) (if feasible). Next, we present stochastic approximation methods to adapt the fairness parameters γ_i in a way that the robustness constraint in (26) is satisfied as well (if feasible) and therefore a solution for HSDPA scheduling problem is obtained. The proposed opportunistic algorithm is denoted by *Algorithm 1* and is summarized in the following steps.

Algorithm 1: Joint Opportunistic Scheduling and MCM Assignment

Step 0—Initialization

Step 1—Estimating per-TTI capacities: Apply (34) to estimate the maximum achievable rate ρ^i_k for all users $i \in \mathcal{I}$ at time k.

Step 2—Computing a throughput optimal MCM (MCS and multicode) for each user: Apply (37) to calculate the optimal MCM denoted by $\alpha(i, x_k)$ for each user $i \in \mathcal{I}$. The optimal MCM of user i is calculated in a way that the throughput of user i is maximized (in case of scheduling user i).

Step 3—Joint MCM assignment and M-LWWF scheduling: Apply (41) to schedule user i_k at time k. Substitute i_k in Step-2 by applying (42) to assign $\alpha(i_k, x_k)$ as the MCM of the scheduled user i_k.

Step 4—Updating the fairness parameters: Apply Stochastic Approximation algorithm in (46) and (47). Set $k \to k+1$ and go to Step 1.

B. Initialization

We start with empty queues: $V^i_0 = 0$, $i \in \mathcal{I}$ and assume all the fairness parameters are identical and equal to one: $\gamma^i_0 = 1$, $i \in \mathcal{I}$ (note that later in the simulations, we may also consider random initial fairness parameters).

C. Estimating Per-TTI Capacities

Here, Step-1 of Algorithm 1 is discussed. The main result of this step is to show how to evaluate the per-TTI capacities μ^i_k, $i \in \mathcal{I}$ in the M-LWWF algorithm (since the occurrence of a frame error is not known a priori, the relation in (15) can not be directly used). We prove that instead of the exact values of the

per-TTI capacities μ_k^i, $i \in \mathcal{I}$, the optimal estimate (i.e., conditional expectation) of μ_k^i can be used by the scheduler without degrading the stability of the queues.

Consider the optimal (minimum-variance) estimate of $\mu_k^i, i \in \mathcal{I}$, defined as

$$\tilde{\mu}_k^i := \mathbf{E}\left(\mu_k^i | x_k\right). \tag{29}$$

Now consider the following well-known smoothing property of conditional expectations [14, Th. 10, Ch. 4.6, p.106]:

$$\mathbf{E}\left(\mu_k^i \mathbf{1}_{\{\phi(x_k)=i\}}\right) = \mathbf{E}\left(\mathbf{E}\left(\mu_k^i | x_k\right) \mathbf{1}_{\{\phi(x_k)=i\}}\right).$$

Substituting from (29) gives

$$\mathbf{E}\left(\mu_k^i \mathbf{1}_{\{\phi(x_k)=i\}}\right) = \mathbf{E}\left(\tilde{\mu}_k^i \mathbf{1}_{\{\phi(x_k)=i\}}\right). \tag{30}$$

The stability condition in (25) can then be reformulated as

$$\mathbf{E}\left(\tilde{\mu}_k^i \mathbf{1}_{\{\phi(x_k)=i\}}\right) > D_i. \tag{31}$$

The relations in (30) and (31) state that any scheduling policy that keeps the buffers stable with per-TTI capacities $\tilde{\mu}_k^i, i \in \mathcal{I}$, will do so as well with per-TTI capacities $\mu_k^i, i \in \mathcal{I}$, (and vice versa).

Assuming perfect channel information, $\tilde{\mu}_k^i$ has a deterministic value at time k. From (29) and (15), $\tilde{\mu}_k^i$ is given by

$$\tilde{\mu}_k^i = \mathbf{E}\left(r_k^i \mathbf{1}_{\{\text{No Frame Error}\}} | x_k\right) = r_k^i\left(1 - f_k^i\right), \tag{32}$$

where r_k^i and f_k^i are defined in (6) and (8), respectively. In the proposed algorithm, we use the deterministic value in (32) as the measure for user i capacity at time k which is equivalent to the original definition in (15) in terms of stability of queues. Here, we only need to consider the MCM combinations that satisfy the maximum FER constraints in (27). Define

$$\mathcal{C}^i(x_k) := \left\{(m,n) \in \mathcal{M} \times \mathcal{N} \quad \text{s.t.} \quad 0 < f_k^i \le f_{max}^i\right\}. \tag{33}$$

Let ρ_k^i be the maximum per-TTI capacity (estimate) of user i at time k. We have

$$\begin{aligned}\rho_k^i &= \max_{(m,n)\in\mathcal{C}^i(x_k)} \tilde{\mu}_k^i \\ &= \max_{(m,n)\in\mathcal{C}^i(x_k)} r(m,n)\left[1 - f^i\left(m,n,c_k^i\right)\right],\end{aligned} \tag{34}$$

where $r(\cdot)$ and $f^i(\cdot)$ are defined in (7) and (8), respectively. Also define

$$\rho_k = \left(\rho_k^1, \rho_k^2, \ldots, \rho_k^L\right). \tag{35}$$

Note that in practice, the exact form of $f^i(\cdot)$ due to the use of Turbo-code and complexity of the system may not be known. Later in the simulations, we use the simulation results and measurements from Motorola research group to map the channel state to the corresponding FER [10].

D. Computing Optimal MCMs

Here we elaborate step 2 of Algorithm 1. We compute the *throughput optimal* MCM for each user $i \in \mathcal{I}$ at time k, i.e., the MCM that gives the highest throughput for user i (in case of scheduling user i at time k). Define

$$\alpha(i, x_k) = \left(m_k^i, n_k^i\right), \quad \phi : \mathcal{X} \times \mathcal{I} \to \mathcal{M} \times \mathcal{N}, \tag{36}$$

where m_k^i and n_k^i are given by (4) and (5), respectively. In order to satisfy (31), the pair (m_k^i, n_k^i) must be chosen to maximize the per-TTI capacity (estimate) $\tilde{\mu}_k^i$ while keeping the instantaneous FER below the allowable threshold. Therefore, by using (32) and (8) the following gives the optimal MCM for each user:

$$\alpha(i, x_k) = \max_{(m,n)\in\mathcal{C}^i(x_k)} \tilde{\mu}_k^i \tag{37}$$

$$= \max_{(m,n)\in\mathcal{C}^i(x_k)} r(m,n)\left[1 - f^i\left(m,n,c_k^i\right)\right] \tag{38}$$

where $\mathcal{C}^i(x_k)$ is defined in (33).

E. Joint M-LWWF User Scheduling and MCM Assignment

Here, Step-3 of algorithm 1 is elaborated. First consider a joint scheduling and MCM assignment policy Λ as defined as in (18). We provide additional degrees of freedom if we introduce a parameter vector γ in the policy Λ and define

$$\Lambda(\cdot, \gamma) = \{\phi(\cdot, \gamma), \psi(\cdot, \gamma)\}, \quad \Lambda : \mathcal{X} \to \mathcal{I} \times \mathcal{M} \times \mathcal{N}.$$

Here, $\phi(\cdot, \gamma)$ is the scheduling policy, $\psi(\cdot, \gamma)$ is the MCM assignment rule, both depending on γ, and

$$\gamma = (\gamma_1 \gamma_1 \ldots \gamma_L), \tag{39}$$

where $\gamma_i \in \mathcal{R}_+, i \in \mathcal{I}$ are *any* arbitrary positive real constants denoted by *fairness parameters* (parameters of the algorithm).

The main result of this section can be stated by the following theorem which presents a parametric joint scheduling and MCM assignment policy that guarantees the stability of queues [see the constraint in (25)].

Theorem 1: Assume maximum per-TTI capacity ρ_k, defined in (35), is an irreducible finite-state discrete-time Markov process. Consider the following parametric policy:

$$\Lambda(\cdot, \gamma) = \{\phi(\cdot, \gamma), \psi(\cdot, \gamma)\}, \quad \text{where}: \tag{40}$$

$$\phi(x_k, \gamma) = \arg\max_{i\in\mathcal{I}} \gamma_i V_k^i \rho_k^i, \quad (\text{M} - \text{LWWF user scheduling}) \tag{41}$$

$$i_k := \phi(x_k, \gamma), \quad (\text{Scheduled user index})$$

$$\psi(x_k, \gamma) = \alpha(i_k, x_k), \quad (\text{MCM assignment}) \tag{42}$$

where i, $i_k \in \mathcal{I}$, $x_k \in \mathcal{X}$, and $k \in \mathcal{Z}_+$. Then for any set of parameters $\gamma \in \mathcal{R}_+^L$, the policy $\Lambda(\cdot, \gamma)$ satisfies the stability QoS constraints in (25), if it is possible with any other policy.

Recall that $\alpha(\cdot)$ is defined in (37), ρ_k^i and V_k^i are defined in (34) and (9), respectively. Distribution parameters $\gamma_i \in \mathcal{R}_+$, are *any* arbitrary positive real constants and $\gamma = [\gamma_i] \in \mathcal{R}_+^L$ as in (39).

Proof: As given by (34), ρ_k^i, $i \in \mathcal{I}$, is the estimate of the maximum bit rate available to user i at time k. With the understanding that (31) is equivalent to (25), the scheduling policy described in (41), will be reduced to the M-LWWF scheduling algorithm introduced in [8]. Additionally, since by assumption the vector ρ_k is an irreducible discrete-time Markov process, the proof in [8] is directly applicable. ∎

Remark: Note that the Markovian assumption on ρ_k in Theorem 1 is not restrictive. In fact, if we consider a one-to-one map from the Markovian channel state c_k (or x_k) to ρ_k then the assumption of Theorem 1 holds.

The user scheduling algorithm in (41) simply selects the user for which the weighted product of the queue content and the channel quality is maximum. This way users with larger queue contents, larger fairness parameters and better channel qualities (compared to the other users) receive higher scheduling priority compared to the other users.

Remark on Feasibility: Theorem 1 states that if stability is feasible then the proposed algorithm will make the queues stable. In this view, the proposed algorithm can be used as a feasibility test for the queues stability. Numerical methods [15] can be used to verify the stability or boundedness of the queues within a desired confidence interval.

F. Updating Fairness Parameters by Stochastic Approximation

Here we elaborate Step-4 of Algorithm 1. The main result of this step is to update the fairness parameters by Stochastic Approximation as in (46) and (47).

Assuming the problem is feasible, Theorem 1 states that the policy $\Lambda(\cdot,\gamma)$, defined in (40), satisfies the stability QoS in (25) for all $\gamma = [\gamma_i] \in \mathcal{R}_+^L$. Therefore, there exists a stationary regime for each queue content process. We can therefore rewrite the robustness QoS in (24) by dropping the time index k

$$\mathbf{P}(V^i > \eta_i) \leq \delta_i, \quad i \in \mathcal{I}. \tag{43}$$

We employ the method of stochastic approximation to adjust the parameter vector γ so that the policy $\Lambda(\cdot,\gamma)$ satisfies the robustness constraint in (43) as well and hence a solution to the HSDPA scheduling problem is obtained. Let $\mathcal{H} \subseteq \mathcal{R}_+^L$ be the feasible region of the policy $\Lambda(\cdot,\gamma)$ in terms of parameter vector γ. We propose a stochastic approximation algorithm to generate a sequence of updates $\gamma_k = (\gamma_k^1 \gamma_k^2 \ldots \gamma_k^L)$ at time $k = \{0,1,2,\ldots\}$ which converge to a feasible parameter vector $\gamma^* \in \mathcal{H}$. Note that with each γ_k, a policy $\Lambda(\cdot,\gamma_k)$ is associated. The initial value γ_0 is set to the vector of ones $\gamma_0^i = 1$, $i \in \mathcal{I}$ which implies that initially no additional priority is granted to any user.

Consider the robustness QoS constraint in (43). In general, the complementary distribution of the buffer content $\mathbf{P}(V^i > \eta_i)$ is not known. The stochastic approximation algorithm estimates $\mathbf{P}(V^i > \eta_i)$ at each time slot and then uses this estimate to adjust the parameter γ_k^i at time k [16]. The procedure is formulated as follows:

Algorithm: Assume the policy $\Lambda(\cdot,\gamma_k)$, defined in (40), is applied at time k. Consider the event $\{V_s^i > \eta_i\}$, where $i \in \mathcal{I}$ denotes the user number and s denotes time. An unbiased estimate of the overflow probability $\mathbf{P}(V^i > \eta_i)$ can be obtained at time k by the number of times that the event $\{V_s^i > \eta_i\}$ occurs up to time k divided by k. Let $\mathcal{P}_k^i$ denote the unbiased estimate of $\mathbf{P}(V^i > \eta_i)$ at time k. We have $\mathcal{P}_k^i = (1/k)\sum_{s=1}^k \mathbf{1}_{\{V_s^i > \eta_i\}}$. Writing by iteration gives

$$\mathcal{P}_{k+1}^i = \left(\frac{k}{k+1}\right)\mathcal{P}_k^i + \left(\frac{1}{k+1}\right)\mathbf{1}_{\{V_{k+1}^i > \eta_i\}}. \tag{44}$$

If the estimate $\mathcal{P}_k^i$ is exceeding the threshold δ_i, then it means that user i, based on the current estimate, does not receive his/her required QoS. Hence, the parameter γ_k^i is increased proportionally to assign higher priority (higher service share) to user i. If the estimate $\mathcal{P}_k^i$ does not exceed the threshold δ_i, γ_k^i is left unchanged

$$\gamma_{k+1}^i = \gamma_k^i + a_k\left(\mathcal{P}_k^i - \delta_i\right)\mathbf{1}_{\{\mathcal{P}_k^i > \delta_i\}} \tag{45}$$

where the step size can be set to $a_k = 1/(k+1)$ for the stationary case i.e., when the true parameter is constant. In the nonstationary case (when the true parameter is slowly varying with time), a_k can be set to a small constant to track the slow variations of the system [5], [16]. Assuming stationarity and $a_k = 1/(k+1)$, (44) and (45) are collectively written as

$$\mathcal{P}_{k+1}^i = \mathcal{P}_k^i + a_k\left(\mathbf{1}_{\{V_{k+1}^i > \eta_i\}} - \mathcal{P}_k^i\right) \tag{46}$$

$$\gamma_{k+1}^i = \gamma_k^i + a_k\left(\mathcal{P}_k^i - \delta_i\right)\mathbf{1}_{\{\mathcal{P}_k^i > \delta_i\}} \tag{47}$$

with initial values of $\gamma_0^i = 1$ and $\mathcal{P}_0^i = 0$. Equations (46) and (47) can be written collectively in vector format as follows:

$$\mathcal{P}_{k+1} = \mathcal{P}_k + a_k\left(\vec{\mathbf{1}}_{\{V_{k+1} > \eta\}} - \mathcal{P}_k\right) \tag{48}$$

$$\gamma_{k+1} = \gamma_k + a_k(\mathcal{P}_k - \delta)\mathrm{diag}\left\{\mathbf{1}_{\{\mathcal{P}_k > \delta\}}\right\} \tag{49}$$

where $\eta := (\eta_1,\ldots,\eta_L)$, $\delta := (\delta_1,\ldots,\delta_L)$, $\mathcal{P}_k := (\mathcal{P}_k^1,\ldots,\mathcal{P}_k^L)$, and $\vec{\mathbf{1}}_{\{V_k > \eta\}} := (\mathbf{1}_{\{V_k^i > \eta_1\}},\ldots,\mathbf{1}_{\{V_k^L > \eta_L\}})$ with initial values of $\gamma_0 = \vec{1}$ and $\mathcal{P}_0 = \vec{0}$ ($\vec{1}$ denotes a $(1 \times L)$ row vector with all elements equal to one). Also note that $\mathrm{diag}\{\vec{y}\}$ represents a diagonal matrix whose diagonal elements are the elements of $\vec{y}$. The algorithm proceeds by applying the policy $\Lambda(\cdot,\gamma_{k+1})$ at time $k+1$. The iterations (48) and (49) will be repeated by setting $k \to k+1$ until the convergence is obtained (assuming a feasible policy $\Lambda(\cdot,\gamma^*)$ exists). Starting with the policy $\Lambda(\cdot,\vec{1})$, numerical results in Section V show that if a feasible $\Lambda(\cdot,\gamma^*)$ exists then the algorithm in (46) and (47) converges rapidly to a feasible point in $\mathcal{H}$.

G. Convergence of the Stochastic Approximation Algorithm

Here we discuss the convergence of the stochastic approximation algorithm in (46) and (47) (with decreasing step size).

Discussion: Since the state of the system $x_k = (v_k, c_k)$ is markovian, we need to use the Ordinary Differential Equation (ODE) approach of [16] to prove the convergence of the Stochastic Approximation algorithm in (46) and (47). The main idea behind the ODE approach is that the asymptotic behavior of (46) and (47) is captured by a deterministic ordinary differential equation which represent the mean trajectory of the system

[16]. This ODE is obtained by stochastic averaging of (46) and (47) and reads [16]

$$\frac{d\mathcal{P}_t^i(\gamma_t)}{dt} = \mathbf{P}\left(V^i(\gamma_t) > \eta_i\right) - \mathcal{P}_t^i(\gamma_t) \tag{50}$$

$$\frac{d\gamma_t^i}{dt} = \left(\mathcal{P}_t^i(\gamma_t) - \delta_i\right)\mathbf{1}_{\left\{\mathcal{P}_t^i(\gamma_t) > \delta_i\right\}} \tag{51}$$

where $i \in \mathcal{I}$ and the variable t denotes continuous time. Note that for clarity, in the above equations we explicitly denote the dependence of V^i and $\mathcal{P}_t^i$ on γ. If stable, the ODEs in (50) and (51) converge to their fixed points P^* and γ^*, where

$$P^* := \left(P_*^1, \ldots, P_*^L\right), \quad \gamma^* := \left(\gamma_*^1, \ldots, \gamma_*^L\right) \tag{52}$$

$$P_*^i = P\left(V^i(\gamma^*) > \eta_i\right) \leq \delta_i. \tag{53}$$

Note that $\gamma^* \in \mathcal{H}$ is a feasible point of the policy $\Lambda\{\cdot, \gamma\}$.

Theorem 2: Under the condition that P^* and γ^* are bounded with probability 1 (i.e., a feasible solution exists), the estimates $\{\mathcal{P}_k\}$ and $\{\gamma_k\}$, generated by algorithm in (48) and (49), converge with probability 1 to the fixed points P^* and γ^* defined in (52) and (53).

Proof: The convergence proof requires showing the convergence of the Stochastic Approximation algorithm for Markovian state-dependent noise. The proof follows from [16, Th. 4.2, Ch. 8.4 p. 243]. The conditions required for convergence are verified as follows.

i) The averaging condition: (A.4.12) in [16] holds since the pair (V_k, c_k) is a positively Harris recurrent process (see the discussion in [8] for the proof of Harris recurrence).

ii) Uniform integrability: the uniform integrability condition (A.4.11) for the vector

$$\left\{\mathbf{1}_{\left\{V_{k+1}^i > \eta_i\right\}} - \mathcal{P}_k^i, \left(\mathcal{P}_k^i - \delta_i\right)\mathbf{1}_{\left\{\mathcal{P}_k^i > \delta_i\right\}}\right\} \tag{54}$$

holds since $\mathcal{P}_k^i$ is uniformly bounded between 0 and 1. Also note that γ_k can be always kept bounded by projection into a compact set. The proof therefore follows from [16, Th. 4.2]. ■

For situations where statistics of the channel are slowly time varying, the optimal parameter vector γ^* will be time varying. Denote the optimal time-varying parameter γ^* by γ_t^*. In such cases it is necessary to adaptively track the time-varying γ_t^*. This can be done by using a constant step size version of (46) and (47) with $a_k = a$, where a is a small constant. For this adaptive case it can be shown, using [16, Th. 4.1, Sec. 8.4.1, p. 240] that the sequence $\{\mathcal{P}_k^i\}$ and $\{\gamma_k^i\}$ converges weakly (in distribution) to P_*^i and γ_*^i for $i \in \mathcal{I}$.

V. Numerical Results

A summary of simulation parameters is presented in Table II. Furthermore, simulation scenarios are described in Section V-A.

Channel: The channel is first simulated by a Rayleigh slow and flat fading model with a Doppler spectrum (Jake's model [17]). Let $g_i(k)$ be the available SNR for user i at time k. Based on the Jake's model, $g_i(k)$ can be written as the transmitted SNR scaled by a fading component and a path-loss component. Expressing in dB, we have:

TABLE I
Modulation and Coding Schemes

MCS	Modulation	Code Rate	Data Rate (1 Code)	Data Rate (15 Codes)
1	QPSK	$\frac{1}{4}$	0.6 Mbps	1.8 Mbps
2	QPSK	$\frac{1}{2}$	1.2 Mbps	3.6 Mbps
3	QPSK	$\frac{3}{4}$	1.8 Mbps	5.3 Mbps
4	16QAM	$\frac{1}{2}$	477 kbps	7.2 Mbps
5	16QAM	$\frac{3}{4}$	712 kbps	10.8 Mbps

TABLE II
Simulation Parameters

Cell radius R_{cell}	500m
Max. number of users L	15
Duration of time slot ΔT	2 ms
Spreading factor	16
Channel estimation	Perfect
Carrier frequency	2 GHz
Chip rate	3.84 Mcps
f_{max}^i	0.1
Transmission SNR: $\frac{P_t}{P_n}$	122dB
Fast fading model	Rayleigh-fading
Propagation model: $L(R)$	$128.1 + 37.6\log(R)$
Channel model	10 State Markov chain
UE buffer time constant t_B	1 second (for all users)
Simulation duration	150000 ΔT (5 min)

$$g_i(k)_{(dB)} = (P_t/P_n)_{(dB)} - L(R)_{(dB)} + \left\{\alpha_i(k)^2\right\}_{(dB)}. \tag{55}$$

where P_t is the transmission power, P_n is the noise power (includes interference as well), L(R) is the path-loss component in terms of the distance R and $\alpha_i(k)$ is the Rayleigh fading gain process for the user i. Based on common assumptions for HSDPA in [18], we adopt the following path-loss model:

$$L(R)_{(dB)} = 128.1 + 37.6\log_{10}(R) \tag{56}$$

where R is the distance of the user from the BS in Kilometers. Furthermore, in the simulations we choose a value for P_t/P_n to obtain reasonable SNRs for the UE-to-BS distances. We assume:

$$P_t/P_n = 122(dB). \tag{57}$$

At the next step, to justify the assumptions of the paper, we present the simulated Rayleigh channel by a FSMP. Here, we model the simulated fading gain process $\alpha_i(k)$ (or equivalently $g_i(k)$) by a ten-state Markov process $c_i(k)$. In general, $c_i(k)$ can be correlated between different users, however, in our simulations independent $\alpha_i(k)$ (or equivalently $c_i(k)$), $\{i = 1, 2, \ldots,\}$ are considered. Five states are assigned for the SNR region corresponding to the single code transmission and five states are assigned to the SNR region corresponding to 15-code transmission. By examining the SNR-FER curves, obtained by Motorola research group [10], the number of states is chosen in a way to attain sufficient resolution for instantaneous capacities. For example, if for some SNR region the FER

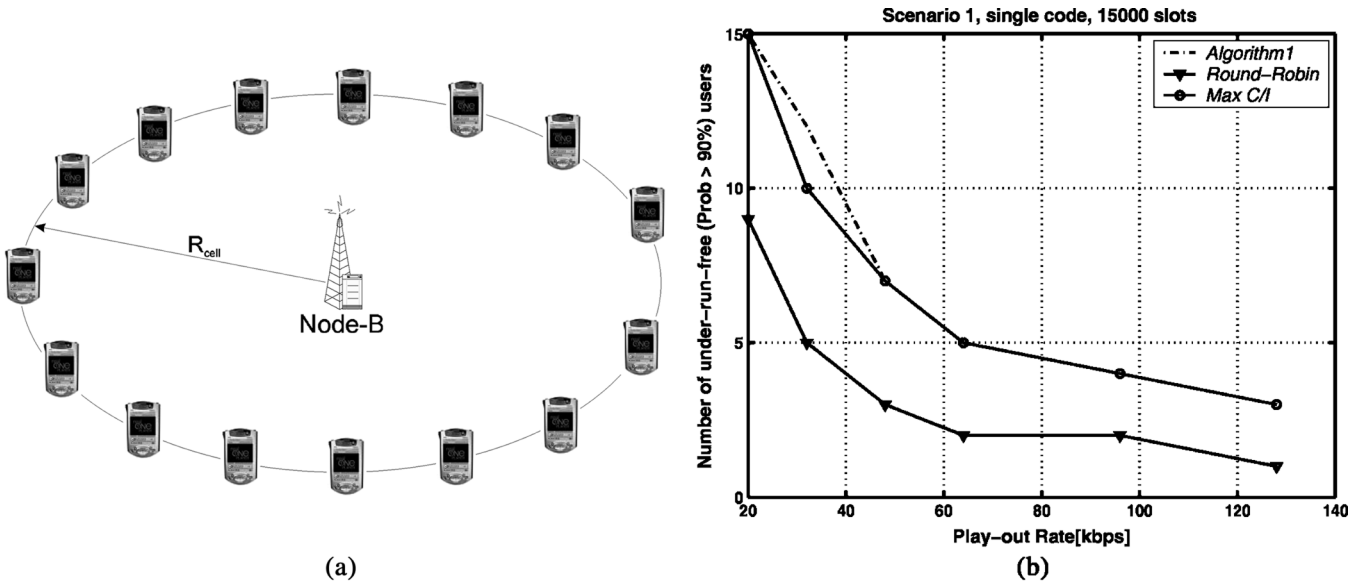


Fig. 2. Comparison of maximum supported HSDPA users in scenario 1. (a) HSDPA users located on the cell edge. (b) Maximum HSDPA users supported versus streaming rates.

curve is more or less constant, only one state is assigned to that region. On the other hand, if FER has large deviations in a particular SNR region, more states are needed to model that SNR region (see for a rigorous procedure for choosing number of states and the threshold levels in [11], and [12]).

MCS and Multicode: In our simulations, the MCS and multicode selection is based on Table I that shows the numerical values for MCS number and spreading codes of the users. These numbers are extracted from the descriptions of HSDPA provided in [2]. For simplicity, we assume either a single code or 15 codes are used for each user. The selected MCS roughly corresponds to the frame error rate of $f_k^i = f_{max}^i = 0.1$ at any time slot.

Parameters: Table II summarizes the parameters and assumption used for our simulation. We have tried to choose realistic parameters from common simulation assumption for HSDPA provided by Ericsson, Motorola, and Nokia in [18]. The simulation starts with full UE buffers.

A. Simulation Scenarios

We compare the performance of our opportunistic policy with the well-known Round-Robin scheduling and Max C/I scheduling. In Round-Robin, scheduling time is shared evenly among the users but the channel or buffer variations are not exploited (nonopportunistic scheduling). In Max C/I, the scheduler simply picks the best channel (highest signal to noise/interference). Therefore, Max C/I can be regarded as a *greedy* opportunistic scheduling.

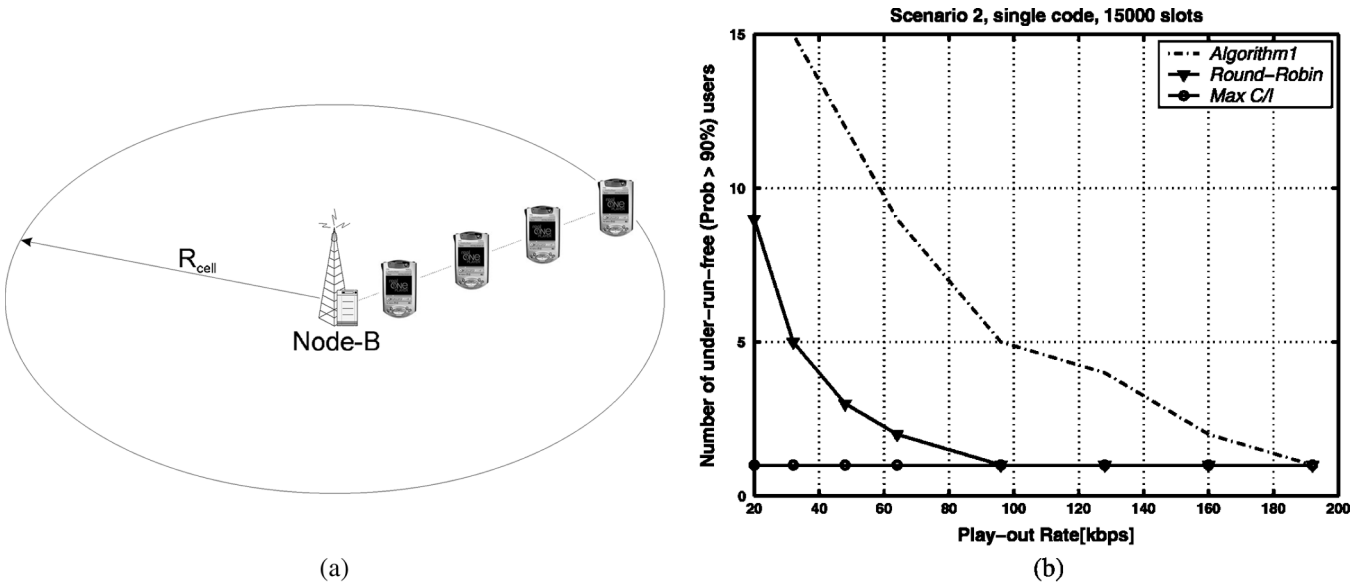
Three scenarios are considered in terms of the distribution of the users in a single wireless cell. For each scenario, we determine the maximum number of the users that can be supported with $p_{success} \geq 90\%$ at a given streaming rate D, where $p_{success}$ is the probability of uninterrupted play-out (assumed to be the same for all users). Therefore, in the QoS formulation in (24), we roughly have: $\mathbf{P}(U_k^i = 0) \leq 0.9$. Hence, $\theta_i = 0$ and $\delta_i = 0.1$ for $i = \{1, 2, \ldots, L\}$. Note that the condition $U_k^i \leq 0$ is equivalent to $U_k^i = 0$ since U_k^i is always non-negative by definition. The simulation scenarios are elaborated in the following. Note that the terms BS and Node-B may be used interchangeably to denote the base-station which is assumed to be located at the cell center.

Scenario 1: The first scenario we consider is a worst case scenario. All users are located on the cell edge [$R = R_{\text{cell}} = 500$ m; see Fig. 2(a)] and consequently have relatively poor channels which only support low data rates (note that in all figures, Node-B denotes the base-station). In this scenario single-code transmission is used.

As shown in Fig. 2(b), in scenario 1, our algorithm (algorithm 1 in Section IV-A) gives the best performance. Max C/I also performs exceptionally well because of the given scenario: all users have the same distance to the BS with the same streaming rate. In this case, it is a *relatively* good solution to only pick the instantaneously best channel without regarding their queue lengths. The total average throughput will be maximized in this case and because of the symmetry, no user will be particularly starved. In the long run, each user receives more or less the same portion of the maximized throughput and hence the overall performance is relatively good. Round-Robin performs poorly in this scenario since channel variations are not exploited."

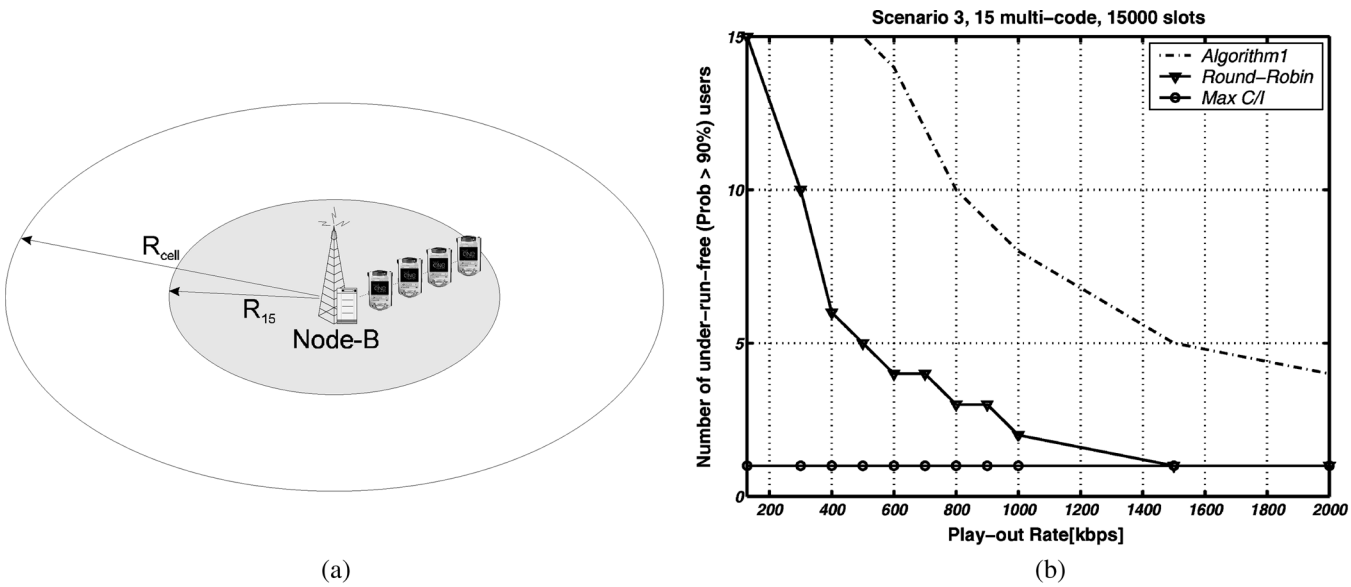
Scenario 2: In this scenario, we assume that the users are distributed uniformly throughout the cell ($10\,\text{m} \leq R \leq 500\,\text{m}$) [see Fig. 3(a)].

The distances of the users and the BS (Node-B) are $R_i = (i/(L-1)) \cdot (R_{\text{cell}} - l_{\min}) + l_{\min}$ where $l_{\min}$ is the minimum distance a UE has from the BS. We observe in Fig. 3(b) that our algorithm (Algorithm 1) performs significantly better than the other two algorithms. Max C/I shows very poor performance because users that are further away from the BS are not served at all. We can see in Fig. 3(b) that Round-Robin gives significantly better performance than Max C/I.



(a) (b)

Fig. 3. Comparison of maximum supported HSDPA users in scenario 2. (a) Users distributed uniformly from the cell edge to the Node-B. (b) Maximum HSDPA users supported versus streaming rates.



(a) (b)

Fig. 4. Comparison of maximum supported HSDPA users in scenario 3. (a) HSDPA users distributed uniformly close to the node-B. (b) Maximum HSDPA users supported versus streaming rates.

Scenario 3: In this scenario, we concentrate our simulation on the cell center ($10\,\mathrm{m} \leq R \leq 250\,\mathrm{m}$), where high data rates can be achieved. We use 15 multicodes for transmission and this way increase our transmission rates significantly.

As shown in Fig. 4(a), the users are distributed uniformly between the BS (Node-B) and the edge of the cell center ($10\,\mathrm{m} \leq R \leq 250\,\mathrm{m}$) and $R_i = (i/(L-1)) \cdot (R_{15} - l_{\min}) + l_{\min}$ where $l_{\min}$ is the minimum UE distance from the BS and R_{15} is the maximum UE distance from the BS (user 15 distance from the center). It is possible to support high streaming rates to a reasonable number of users in the cell center. Again we see in Fig. 4(b) the advantage of our algorithm. The Max C/I algorithm completely fails in this scenario, because users that are further away from the BS are not served at all. We conclude that compared with the Round-Robin and Max C/I scheduling, our proposed algorithm gives a significantly better performance in terms of the maximum number of the users that can be supported with the desired QoS.

VI. Conclusion

In this paper, we propose a discrete event model for streaming users in HSDPA to formulate the QoS requirements of the HSDPA streaming system as a feasibility problem. We then present a practical joint opportunistic user-scheduling and MCM assignment policy as the solution to the above feasibility problem. The proposed policy enables a smooth play-out for HSDPA users if it is possible with any other policy. The numerical results show that employing the proposed opportunistic

policy achieves a significant improvement in the maximum number of users that can be supported with the desired QoS.

ACKNOWLEDGMENT

The authors acknowledge F. Blömer for his help in obtaining the simulation results. They would also like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] F. Blömer, "Scheduling for Streaming Users in HSDPA," M.Sc. thesis, Munich Univ. of Technol., Munich, Germany, Dec. 2003.
[2] T. E. Kolding, F. Frederiksen, and P. E. Mogensen, "Performance aspects of WCDMA systems with High Speed Downlink Packet Access (HSDPA)," in *Proc. IEEE 56th Vehicular Technology Conf.*, Vancouver, BC, Canada, 2002, vol. 1, pp. 477–481.
[3] F. Fredriksen and T. E. Kolding, "Performance and modeling of WCDMA/HSDPA transmission/H-ARQ schemes," in *Proc. 56th IEEE Vehicular Technology Conf. (VTC 2002-Fall)*, Vancouver, BC, Canada, Sep. 2002, vol. 1, pp. 472–476.
[4] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraint in wireless netowork," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 2053–2064, Oct. 2001.
[5] E. C. X. Liu and N. Shroff, "Transmission scheduling for efficient wireless resource utilization with minimum-utility guarantees," in *Proc. IEEE VTC Fall 2001*, Atlantic City, NJ, Oct. 2001, pp. 824–828.
[6] R. J. F. Berggren, S. L. Kim, and J. Zander, "Joint power control and intracell scheduling of DS-CDMA nonreal time data," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 1860–1870, Oct. 2001.
[7] F. Berggren and R. Jänitti, "Asymptotically fair scheduling on fading channels," in *Proc. IEEE 56th Vehicular Technology Conf.*, Vancouver, BC, Canada, 2002, vol. 4, pp. 1934–1938.
[8] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions," in *Bell Labs Tech. Memo.*, 2000.
[9] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.
[10] R. Love, A. Ghosh, R. Nikides, L. Jalloul, M. Cudak, and B. Classon, "High speed downlink packet access performance," in *Proc. 53rd Vehicular Technology Conf. 2001 (VTC 2001)*, May 2001, vol. 3, pp. 2234–2238.
[11] H. S. Wang and N. Moayeri, "Finite-state Markov channel—a useful model for radio communication channels," *IEEE Trans. Vehic. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.
[12] Z. Q. and S. A. Kassam, "Finite-state Markov model for rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.
[13] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2st ed. Belmont, MA: Athena Scientific, 2000.
[14] G. Grimmet and D. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford, U.K.: Oxford, 2001.
[15] G. C. Pflug, *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*, 1st ed. Norwell, MA: Kluwer, 1996.
[16] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.
[17] T. S. Rappaport, *Wireless Communications Principles and Practice*, 1st ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[18] Ericsson, Motorola, and Nokia, "Common HSDPA system simulation assumptions," presented at the 3GPP TSG RAN WG1 Meeting #15, Berlin, Germany, 2000, TSGR1#15(00)1094.

Arsalan Farrokh (S'99-M'00) received the B.S degree in electrical engineering from Sharif University of Technology, Tehran, Iran in 1996 and the M.S degree in electrical engineering from Simon Fraser University, Burnaby, BC, Canada. He is currently pursuing the Ph.D. degree at the University of British Columbia, Vancouver, BC.

His research interests include stochastic optimization with application to multimedia, defense, and business systems.

Vikram Krishnamurthy (S'90-M'91-SM'99-F'05) was born in 1966. He received the Bachelor's degree from the University of Auckland, Auckland, New Zealand, in 1988, and the Ph.D. degree from the Australian National University, Canberra, Australia, in 1992.

Since 2002, he has been a Professor and Canada Research Chair at the Department of Electrical Engineering, University of British Columbia, Vancouver, BC, Canada. Prior to this, he was a Chaired Professor at the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia, where he also served as Deputy Head of the department. His research interests span several areas including ion channels and nanobiology, stochastic scheduling and control, statistical signal processing and wireless telecommunications.

Dr. Krishnamurthy has served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS AEROSPACE AND ELECTRONIC SYSTEMS, the IEEE TRANSACTIONS NANOBIOSCIENCE, SYSTEMS AND CONTROL LETTERS, and the *European Journal of Applied Signal Processing*. He was a Guest Editor of the Special Issue on Bio-Nanotubes of the IEEE TRANSACTIONS ON NANOBIOSCIENCE in March 2005.