Service Differentiation in Multi-Rate HSDPA Systems

Hongxia Sun
Department of Computer Science
University of Calgary
Calgary, AB, Canada T2N 1N4
Email: sunh@cpsc.ucalgary.ca

Carey Williamson
Department of Computer Science
University of Calgary
Calgary, AB, Canada T2N 1N4
Email: carey@cpsc.ucalgary.ca

***Abstract*—In multi-rate cellular transmission systems, users with different Quality of Service (QoS) requirements share the same wireless channel. In this paper, we investigate the problem of efficient resource allocation for scheduling with differentiated QoS support in a multi-rate system. We propose Dynamic Global Proportional Fairness (DGPF) scheduling on the downlink. We investigate the performance of the scheduling algorithm and model the proposed scheme in a High Speed Downlink Packet Access (HSDPA) simulation environment. The simulation results show that our approach can achieve suitable QoS for different classes of users without compromising aggregate network throughput. The results also show that TCP dynamics affect overall system performance.**

I. INTRODUCTION

Commercial cellular data networks are migrating toward a new wireless technology called High-Speed Packet Access (HSPA) [7], [13], [14], [18]. This new technology is being deployed in two phases, with initial support for the downlink (HSDPA), and subsequent support for the uplink (HSUPA). These wireless transmission technologies promise enhanced data rates for mobile Internet users, with peak[1] data rates of up to 14.4 Mbps on the downlink, and up to 5.74 Mbps on the uplink.

The HSDPA and HSUPA technologies are based on Wideband Code Division Multiple Access (WCDMA) for physical layer transmission [18]. CDMA is a clever modulation scheme that allows multiple users to transmit at the same[2] time, yet have receivers correctly extract their intended messages from the intertwined transmissions using pre-assigned orthogonal codes (commonly referred to as Walsh codes in the literature). There are a finite number of Walsh codes available for simultaneous use in the system (e.g., 16), which limits the number of simultaneous user transmissions that can be supported (e.g., 4 at a time, using several codes each). Nonetheless, the concurrent transmission and reception capabilities provide significant throughput advantages over previous bandwidth-constrained cellular data networks.

In these high data-rate systems, multiple users share the same wireless channel, and the channel conditions experienced by different users vary due to their distances from the base stations and the ambient interference. This makes it difficult for the scheduler to allocate resources in a manner that meets the Quality of Service (QoS) requirements for all users. The scheduler's job is even harder because different types of users have different QoS requirements.

A particularly challenging task in such systems is *downlink scheduling*, so that the system can meet the QoS requirements of multiple users, while maintaining high system throughput. In wireless communication systems with a shared medium, a good scheduling policy must balance between three goals: maximizing transmission capacity, satisfying the delay constraints of real-time applications, and achieving fairness amongst users. These are often contradictory aims, and achieving them all is difficult or impossible.

In this paper, we study the problem of efficient resource scheduling for supporting differentiated service in a multi-rate HSDPA system. The proposed scheme considers not only QoS parameters for different classes of users, but also the overall transmission efficiency of the system (i.e., high aggregate system throughput). We assume that deadline-sensitive users have a higher priority than Best-Effort (BE) users. The proposed algorithm dynamically computes the channel resources required for meeting the QoS requirements of the priority users. The remaining system resources are then distributed to the BE users.

Instead of using conventional Proportional Fairness (PF) scheduling [15], we propose a QoS-based global PF algorithm to enhance the transmission efficiency of the system, while dynamically allocating sufficient resources for priority users. We consider general Internet traffic conditions in the system, where the number of users changes with time, and user data transmissions are subject to the end-to-end flow and congestion control policies of TCP (Transmission Control Protocol). We model the schemes in an HSDPA simulation environment to investigate the scheduling impact. Performance advantages are demonstrated by comparing our results to those for a global PF algorithm.

The primary contributions in our work are: (1) the development of a detailed HSDPA system simulator; (2) the design

[1] These are the theoretical peak data rates according to the latest standards. Commercial deployments may not fully achieve these rates, though some providers claim that they are currently achieving up to 7.2 Mbps in commercial HSDPA systems.

[2] Similar principles apply for Orthogonal Frequency Division Multiple Access (OFDMA) systems [4], [5], [9], [10], [11]. We restrict our discussion in this paper to WCDMA and HSDPA.

and evaluation of a novel QoS-based scheduler; and (3) the assessment of TCP effects on HSDPA system performance.

The rest of this paper is organized as follows. Section II provides a brief synopsis of prior related work. Section III presents our system model. Section IV presents our QoS-based scheduling algorithm for two classes of traffic. Section V presents our HSDPA simulation model, and Section VI presents the simulation results. Finally, Section VII concludes the paper.

II. Background and Related Work

Opportunistic scheduling is a widely accepted technique in wireless networks for efficient utilization of scarce radio resources. By allocating transmission capacity to users with good wireless conditions, and deferring transmissions for users with poor wireless conditions, system efficiency is improved. Long-term fairness can be achieved via the inherent time-varying fluctuations of wireless channel conditions, the mobility of users, or explicit mechanisms in the scheduling policy.

Proportional Fairness (PF) scheduling is the most prevalent scheduling algorithm [15]. It provides a trade-off between user fairness and maximizing average system throughput. More recent work [22] has extended PF to more general traffic conditions. For example, in multi-rate systems, a global version of PF scheduling is used to increase throughput and provide fairness by utilizing multi-user diversity to favour users with good channel conditions [16]. However, PF (and its variants) do not provide differentiated QoS support.

Several authors have considered QoS and PF in multi-rate systems [6], [12], [13], [19], [20]. Most of the schemes are heuristic algorithms derived from the PF scheduling for a single carrier system. The algorithms factor QoS constraints into the PF scheduling algorithms using QoS-weighted parameters or pre-defined user utility functions. Although QoS requirements can be met, this actually does not satisfy the PF criteria for multi-user transmission in the system. Therefore, it does not maximize the transmission rate, and subsequently results in lower system throughput.

Many schemes have been proposed for optimizing scheduling or resource allocation in multi-rate systems [4], [10], [11], [17]. The schemes often involve minimizing total transmit power for given rate constraints, or maximizing the weighted sum of rates subject to transmission power constraints. In either case, optimal resource allocation is a difficult combinatorial optimization problem. Simpler heuristics and approximations can be used, as long as they do not compromise the efficiency of the scheduling.

III. System Model

A. Downlink Scheduling in Multi-rate Systems

We consider a multi-user downlink system with packetized transmission. One central Base Station (BS) and multiple geographically distributed users occupy a cell. The BS allocates resources to mobile users via a scheduling algorithm that is based on the channel state information, as well as other possible decision-making criteria.

The basic frame structure of the downlink channel is shown in Figure 1. The horizontal axis is divided into fixed-duration time slots called Transmission Time Intervals (TTIs). The vertical axis represents the number of Walsh codes available in the system. Each user could receive multiple codes in a given time slot. In the wireless environment, adaptive modulation and coding techniques are used to assign appropriate levels depending on channel measurements.

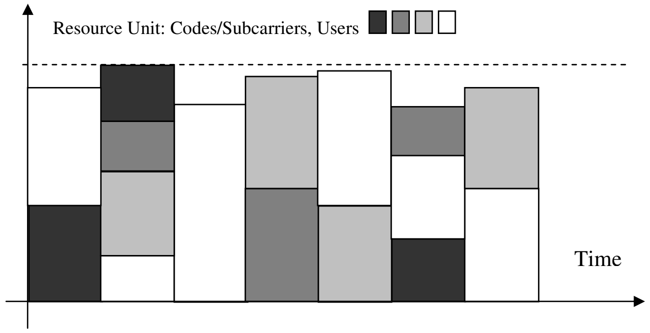


Fig. 1. HSDPA Downlink Scheduling Example

In the example shown in Figure 1, there are four users (Black, Dark Gray, Light Gray, and White) and 15 codes available, as indicated by the dashed line. In the first TTI, two users are chosen for transmission: the Black transmission consumes 6 codes, while the White transmission uses 7 codes, leaving 2 codes idle and wasted. In the second TTI, all 4 users are able to transmit: White (3 codes), Light Gray (5 codes), Dark Gray (3 codes), and Black (4 codes), fully utilizing the system. In the third TTI, the White user has excellent channel conditions, and transmits a high-data-rate burst using 12 codes. In the next TTI, two users transmit with 7 codes each, and so on. The dynamic selection of users for transmissions based on rate requests, channel conditions, and traffic load is determined by the downlink scheduling.

B. Stochastic Traffic and QoS Criteria

The stochastic characteristics of different classes of traffic are determined by the service types (e.g., Web browsing, Voice-over-IP, video streaming). We model traffic flow as an On/Off process representing packet arrivals within a connection, and idle periods within a user session. In particular, we use an HTTP traffic model in our simulation.

We consider two classes of users: delay-sensitive users (Class A), and Best-Effort (BE) users (Class B). The instantaneous number of users in the system varies with time; this value depends on the channel state, resource allocation, and TCP protocol performance.

Packet delay in an HSDPA system is a function of the packet arrival process and the average transmission rate. For a delay-sensitive user, the packet delay constraint implies a minimum average required service rate. QoS guarantees are expressed statistically. For example, the observed packet delay should be less than the delay threshold D with probability q, where D and q are specified QoS parameter settings. The observed distribution of delay depends on the joint distribution of the channel state and the resource allocation. BE users do not have any deadline for their packets. Hence, their rate allocation is flexible and any remaining resources may be used by BE users.

IV. Scheduling Schemes

A. Assumptions and Notation

We model an HSDPA downlink to study the scheduling algorithm and the impact of different classes of traffic in the system. There are several assumptions in our system. First, the channel is a single broadband link shared by all users in a cell. Data transmissions to each user occur in time slots assigned by the scheduler. Second, adaptive modulation and coding schemes are used to support different data rates with reliable transmission. A set of data rates is available based on the estimated channel condition from user feedback. The transmission formats and rates are chosen based on system specifications [1], [14].

The downlink channel is modeled as a discrete time system. Let ψ be the set of active users at time slot t, where $t \in (1, 2, \ldots)$. As users are accepted into the system, or leave the system, the number of active users changes with time.

Table I summarizes the notation used in our model of the scheduling problem. The scheduler selects up to N users among all active users in every slot based on the rules determined by the scheduling algorithm, the QoS requirements, and the resource constraints. There is a separate buffer for pending (queued) packets for each user. At the i-th time slot, the scheduler selects the appropriate N buffers to service, while considering overall system efficiency. The QoS requirement is met by dynamically adjusting assigned resource units for priority users while utilizing global PF scheduling for transmission efficiency and user fairness.

B. Global Proportional Fairness (GPF)

We start with a description of Global PF (GPF) scheduling from the literature [16]. In the i-th time slot, a GPF scheduler selects the appropriate N users for service. GPF maximizes overall system throughput by finding a solution for the optimization problem:

$$\Theta = argMax \prod_{i \in \psi} (1 + \frac{\sum_{k \in C_i} r_{i,k}}{(w-1) \cdot E[Th_i]}) \tag{1}$$

subject to

$$\sum_{i \in U} \sum_{k \in C_i} K_{i,k}^U \leq L \tag{2}$$

where C_i is the set of resource units allocated to user i, $r_{i,k}$ is the k-th instantaneous transmission rate with $K_{i,k}^U$ units in set C_i, L is the total number of resource units in the system, and $E[Th_i]$ is the mean moving average throughput of user i at time slot t over a window size of w slots:

$$Th_i(t) = (t-1)Th_i \cdot (1 - \frac{1}{w}) + \frac{\sum_{k \in C_i} r_{i,k}}{w} \tag{3}$$

C. Dynamic Global PF

To adapt GPF for dynamic On/Off traffic, it is necessary to exclude idle periods of users from the throughput calculation. In our study, the moving average throughput of user i is calculated based on user connections in a set M_i, where $m \in M_i$. The mean throughput is the average throughput of multiple active connections:

$$E[Th_i] = \sum_{m \in M_i} Th_i^{(m)}(t) \tag{4}$$

Our new scheme is called Dynamic Global Proportional Fairness (DGPF). It has two layers in the scheduling architecture. The lower layer adaptively determines the minimum resources required to meet the QoS requirements. The upper layer focuses on the overall transmission efficiency in the system, and global fairness to the users determined eligible by the lower layer. This function is carried out by using a variant of PF scheduling, with multiple concurrent user transmissions in each time slot.

The packet call (TCP connection) is the basic unit for resource allocation for users. The total number of active users in every slot varies with time according to stochastic traffic characteristics and the link transmissions. We divide users into two classes, with class A subject to the QoS constraint, and class B for BE users. Set ψ_A represents class A users in ψ and set ψ_B is for class B users.

Denote $U \subseteq \psi_A$ to be a subset of class A users, and $V \subseteq \psi_B$ to be a subset of class B users selected from ψ. Our scheduling algorithm is designed to optimize the overall throughput of the downlink, but with fairness to users. At the same time, the algorithm considers the QoS service levels for different user classes. The resulting scheduling algorithm finds an allocation Θ of slots at time t that optimizes:

$$\begin{aligned}\Theta = argMax\{ &\prod_{i \in U} (1 + \frac{\sum_{m \in M_i} \sum_{k \in C_i} r_{i,k,m}}{(w-1) \cdot E[Th_i]}) \\ &\cdot \prod_{j \in V} (1 + \frac{\sum_{m \in M_j} \sum_{l \in C_j} r_{j,l,m}}{(w-1) \cdot E[Th_j]})\}\end{aligned} \tag{5}$$

subject to

$$\sum_{i \in U} \sum_{k \in C_i} K_{i,k}^U \leq L_A \tag{6}$$

$$\sum_{i \in U} \sum_{k \in C_i} K_{i,k}^U + \sum_{j \in V} \sum_{l \in C_j} K_{j,l}^V \leq L \tag{7}$$

where C_i is the set of resource units assigned to users i in set U and C_j is the set of resource units to users j in set V. We do not explicitly consider transmission power constraints, since the rates assigned to users must be feasible combinations based on power limitations and channel states. L_A is the dynamic (i.e., time-varying) resource limit for Class A users. L_A has to be controlled and adjusted carefully to meet the user QoS requirements with as few resources as possible.

TABLE I
SUMMARY OF NOTATION USED IN HSDPA SYSTEM MODEL

Symbol	Description
ψ	Set of active users
N	Number of users selected for service
L	Total number of codes available in the system
L_A	Number of codes for Class A users (time-varying)
U	Set of Class A users selected for service
V	Set of Class B users selected for service
U_A	Set of feasible solutions for Class A users
U_B	Set of feasible solutions for Class B users
C_i	Set of channel codes allocated to user i
$r_{i,k,m}$	k-th rate assigned for connection m of user i
$K_{i,k}^U$	Actual resource allocated to k-th rate of user i in U
w	Time window (in slots) used for throughput averaging
Th_i	Throughput for user i (moving average)

D. General Algorithm

In our QoS scheduling, the following steps occur:

- **Step 1**: Initialize solution sets and variables.
- **Step 2**: Set limits for dynamic control policy.
- **Step 3**: Identify feasible combinations of users.
- **Step 4**: Choose best combination of users from the feasible combinations.
- **Step 5**: Find best solution for scheduling.
- **Step 6**: Update state information for next time slot.

E. Detailed Algorithm

The following algorithm determines the best assignment of channel codes to active users in the system, according to the QoS scheduling, along with dynamic adjustment of the resource limit L_A. We assume that there is a group of active users belonging to sets ψ_A and ψ_B at the current scheduling slot. U_A is a set of user sets, with each element representing a feasible solution for the current selected group of users. A user i can be assigned $K_{i,k}$ codes with rate $r_{i,k,m}$ for its m-th connection. The difference ΔT between actual mean throughput $\overline{Th}$ and the expected throughput Th is a variable that controls the dynamic adjustment of resource limit parameter L_A.

The following algorithm assigns codes to users:

```
• Step 1: Initialization
  Set $\Delta T = 0, L_A = \frac{L}{2}, \Delta L = 1, R = 0$.
  Set $U_A := \phi, U_B := \phi$
• Step 2: Outer control loop
  If $\Delta T > 0$, then $L_A = L_A + \Delta L$.
  If $\Delta T < 0$, then $L_A = L_A - \Delta L$.
  Set $U := \phi, V := \phi$
• Step 3: Inner search loop
  $\forall U \subseteq \psi_A$ and $U \not\subseteq U_A$, $U_i \in U$ , $0 < i \leq N_A$,
  and $\sum_{i=1}^{N_A} \sum_{m \in M_i} K_{i,k}^U \leq L_A$
      $U_A := U_A \cup U$
  $\forall V \subseteq \psi_B$ and $V \not\subseteq U_B$, $V_i \in V$, $0 < i \leq N_B$,
  and $\sum_{i=1}^{N_B} \sum_{m \in M_i} K_{i,k}^V \leq L - \sum_{i=1}^{N_A} \sum_{m \in M_i} K_{i,k}^U$
      $U_B := U_B \cup V$
  If $U = \phi$ and $V = \phi$, then Stop.
• Step 4: Quantify the scheduling combinations
  Set $R_i^U = 0$, $R_i^V = 0$, $T_i^U = 0$, $T_i^V = 0$.
  for $i = 1$ to $N_A$
      $T_i^U = T_i^U + \sum_{m \in M_i} \sum_{j \in C_i} r_{i,j,m}$
      $R_i^U = R_i^U + \sum_{m \in M_i} \sum_{j \in C_i} \log(r_{i,j,m})$
  end for $i$
  for $i = 1$ to $N_B$
      $T_i^V = T_i^V + \sum_{m \in M_i} \sum_{j \in C_i} r_{i,j,m}$
      $R_i^V = R_i^V + \sum_{m \in M_i} \sum_{j \in C_i} \log(r_{i,j,m})$
  end for $i$
• Step 5: Choose best scheduling decision
  Set $N := \phi$
  If $R_i^U + R_i^V > R$
  then $R := R_i^U + R_i^V$ and $N := U \cup V$
  Go to Step 2.
• Step 6: Update state
  $\overline{Th}_i(t) = (t-1)\overline{Th}_i \cdot (1 - \frac{1}{w}) + \frac{T_i^U}{w}$,
  $\Delta T = Th - \overline{Th}$
  $t = t + 1$
  Go to Step 1.
```

F. Example

We present a small example to illustrate the workings of the foregoing algorithm. Consider a set of five users with packets pending in the system for downlink transmission. Suppose that users U_1, U_2, and U_3 are Class A users, while users V_4 and V_5 are Class B users. U_1 has good channel conditions, and wants to use a transmission rate that requires 5 Walsh codes. U_2 has poor channel conditions, and wants to use a rate with 2 codes. The transmission rate conditions for the other users are $U_3 = 7$ codes, $V_4 = 10$ codes, and $V_5 = 6$ codes. Suppose that there are a total of 15 codes available in each TTI, and that the resource allocation for Class A users is $L_A = 10$.

In the first TTI, there are many possible choices for feasible transmission schedules for Class A users: U_1 alone, U_2 alone, U_3 alone, U_1 and U_2 together, or U_2 and U_3 together. Note that U_1 and U_3 together is not feasible, since this would exceed L_A. Suppose that the scheduler chooses U_2 and U_3. This choice leaves 6 codes remaining, so user V_5 can be selected. The first TTI thus contains three transmissions (U_2, U_3, and V_5), using 13 codes and wasting 2. Assuming no further packet arrivals, the next TTI can accommodate both U_1 (5 codes within Class A) and V_4 (the remaining 10 codes).

Although the computational complexity of DGPF is high, our scheduling algorithm makes the decisions required for packing user transmissions into the codes available in each TTI. The intent is to maximize system throughput, while also respecting the QoS and fairness requirements of each traffic flow.

V. Simulation Methodology

A. Simulator Design

To investigate performance and evaluate the proposed QoS scheduling, we simulate the scheduling scheme in a HSDPA system. Multiple users are allowed to transmit data in the same TTI, by appropriately sharing the finite channel codes.

We developed our simulator using C/C++, combined with MATLAB to simulate the wireless channel. The simulator architecture is shown in Figure 2.

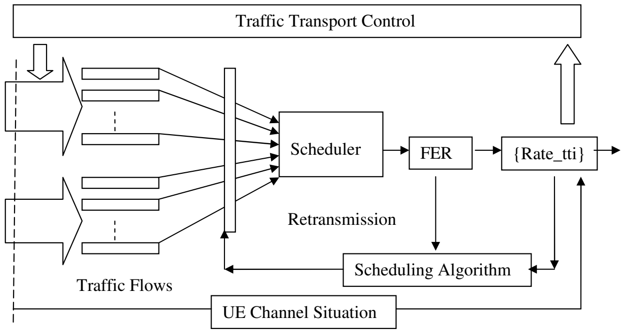


Fig. 2. Simulator Architecture

B. Wireless Network Model

Our simulation model represents a cellular base station situated in the center of 6 neighbouring base stations, which are in turn surrounded by a perimeter of 12 base stations. Each base station has three sectors. We model intra-cell and inter-cell interference in this network, using a multi-path fading wireless channel model. The interference from the 18 surrounding cells is calculated using simulated signal-to-interference-and-noise ratio (SINR). The main simulation parameters are summarized in Table II.

TABLE II
Simulation Parameters

Item	Value
Cellular layout	19 cells
Cell-to-Cell distance	3 km
Path loss exponent	4
Base station power	42 dB
Average noise power	3 dB
StdDev of noise	6 dB
User mobility speed	3 km/hour

We model High Speed Downlink Shared Channel (HS-DSCH) in our simulation to simulate downlink data based on the PHY layer structure. A High Speed Physical Downlink Shared Channel corresponds to one channelization code of fixed spreading factor $SF = 16$ from the set of channelization codes reserved for the HS-DSCH transmission. Each user transmission can be assigned multiple channelization codes in the same HS-PDSCH subframe. The TTI is 2 ms.

The frame error rate (FER) on the link depends on modulation, coding, and retransmission schemes. In our simulation, we simply assume a fixed FER of 2% for all transmissions. Hybrid-ARQ is adopted for quick wireless retransmission. Packet latency on the downlink consists of the queueing delay at scheduler, and the transmission latency over the air link (i.e., frame size divided by transmission rate).

The specific transmission rate formats that we use in our simulator are shown in Table III. When each packet arrives to the downlink transmission queue of the corresponding access terminal, the transmission format is determined by the scheduling algorithm based on the rate table [1], [3].

TABLE III
HSDPA Transmission Rate Table

SINR Value (dB)	Transmission Rate (Kbps)	Walsh Codes	Frame Size (bits)
-2.81	68	1	137
-2.37	86	1	173
-1.66	116	1	233
-0.72	158	1	317
-0.07	188	1	377
0.79	230	1	461
2.55	325	2	650
3.72	396	2	792
4.75	465	2	931
6.78	631	3	1,262
7.83	741	3	1,483
8.81	871	3	1,742
10.11	1,140	4	2,279
10.53	1,292	4	2,583
11.13	1,660	5	3,319
11.33	1,782	5	3,565
12.18	2,095	5	4,189
13.51	2,332	5	4,664
14.53	2,643	5	5,287
15.50	2,944	5	5,887
16.59	3,277	5	6,554
17.59	3,584	5	7,168
18.50	4,860	7	9,719
19.50	5,709	8	11,418
20.50	7,206	10	14,411

C. User Model

In each run of the simulation, a specified number of users (from 40 to 200) are randomly placed into the network with uniform geographic distribution. We simulate a scenario in which users download data on a HSDPA link from the base station using TCP connections. For Web traffic users, 80% are static during their sessions and 20% are mobile. Each mobile user moves according to the waypoint mobility model [8]. A pedestrian movement speed of 3 km/h is assumed. We do not simulate handoff in this study, but simply assume conservation of users across sector boundaries.

Our experiments simulate 2 hours of the transmission link with at least 500,000 packets transmitted through the scheduler.

D. Application-Layer Traffic Model

The simulator has a realistic Web browsing traffic model, following the HTTP model proposed in [2]. In the model, a typical Web browsing session is divided into On/Off periods representing Web page downloads and the intermediate reading times.

This hierarchical traffic model has four levels: *session*, *packet-call*, *object*, and *packet*. The session level represents a user Web browsing session, which may last for several minutes or more. The packet-call level represents a complete Web page, as obtained from a Web server using a single persistent TCP connection. The object level represents the individual files that are retrieved as part of the complete Web page. The packet level models individual TCP/IP packets for each object. Packets are usually 1500 bytes in size. Each packet typically requires between 2-10 milliseconds to transmit on the HSDPA channel, depending on the wireless channel conditions and the transmission rate selected.

The packet traffic characteristics within a packet call depend on the version of HTTP used by the Web servers and browsers. Currently two versions of the protocol, HTTP/1.0 and HTTP/1.1, are widely used by the servers and browsers. These two versions differ in how the transport layer TCP connections are used for the transfer of the main object and the embedded objects (if any). We use the HTTP/1.1 persistent connection model in our simulation.

E. TCP Model

TCP on the Internet uses a feedback-based congestion control mechanism, wherein positive acknowledgements from a receiver advance not only the sliding window used for flow control, but also expand the congestion window size ($cwnd$) used for adaptive congestion control. The 3GPP2 document [2] recommends a TCP-like approximation to this behaviour, wherein two new data packets are launched for each acknowledgement received. This model reflects TCP's self-clocking behaviour, as well as the $cwnd$ expansion process, although it does not actually model TCP timeout, fast retransmit, or fast recovery following a packet loss.

We use persistent TCP connections in our simulation model. With persistent connections, successive objects (files) within the same Web page are transferred using the TCP congestion window size resulting from the preceding transfer (rather than starting over with a default TCP window size of one packet), and are thus able to achieve much higher throughput. One of the side effects, however, is very bursty traffic, since prolonged Web sessions can generate large bursts of data packets in the system (see Figure 3).

F. Scheduling and Service Classes

Our HSDPA simulator supports two traffic classes: data (HTTP) and voice (VoIP). In this paper, we only consider HTTP traffic. Our prior work has studied the interactions between HTTP and VoIP traffic [21].

Three different scheduling algorithms are modeled in the simulator. The first one is Proportional Fairness (PF). This scheduler is the *de facto* standard in cellular networks research, providing high throughput and reasonable fairness (but no QoS) for network users. We use it in validation experiments (not shown here), and as a baseline comparison in our work. The second scheduler is called Global PF. The classic PF algorithm schedules a *single* user (with the highest PF value) in a given time slot in a CDMA system. In HSDPA, *multiple* users can be scheduled in the same TTI, as long as codes are available. The Global PF algorithm extends PF to this case, and typically achieves much higher throughput than PF. The third algorithm is a QoS-based scheduler. This algorithm considers the service requirements (e.g., delay, throughput) of different traffic classes, and dynamically partitions network resource usage to achieve a target operating point that appropriately balances the demands of the two traffic classes. This algorithm is effective in controlling the relative throughput allocations among traffic classes. However, the aggregate throughput is often lower than that achieved by PF.

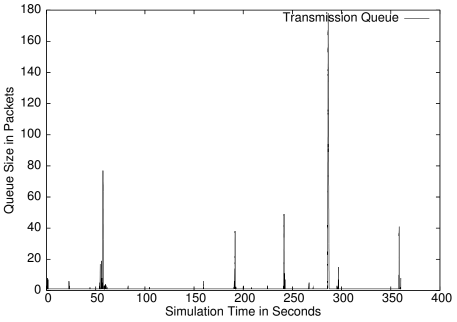


Fig. 3. Example of HTTP/TCP Traffic Model (1 User)

G. Performance Metrics

We study the user-perceived throughput, rather than the network-level throughput that the channel is able to achieve. The former relates to the real traffic in the system and the latter relates to the transmission capacity of the channel. In the following simulation results, all throughputs indicate user-perceived throughput for each class of users, based on the specified scheduling algorithm.

VI. Simulation Results

This section presents simulation results demonstrating the capabilities of our HSDPA network simulator, and the performance of the DGPF scheduler.

A. Comparison to GPF

Figure 4 provides a throughput comparison between the proposed DGPF scheduler and the Global PF scheme. From Figure 4, it can be seen that the aggregate user-perceived throughputs for the two scheduling algorithms are similar (around 1.2 Mbps). However, the Class A priority users obtain higher mean service rate than the Class B users with DGPF scheduling, while both Class A and Class B receive the same service with GPF scheduling. More importantly,

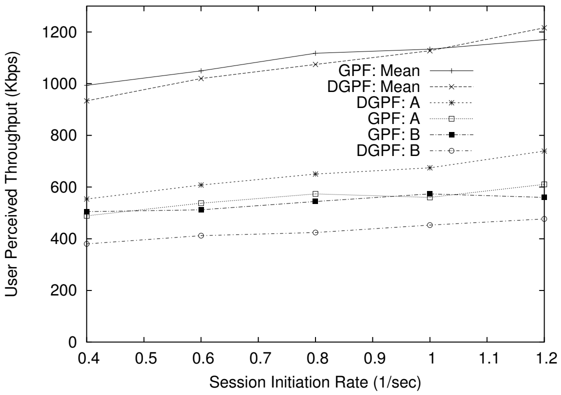


Fig. 4. Comparison of DGPF and GPF

the DGPF scheduling provides service differentiation without compromising the overall system throughput.

The results in Figure 4 show that the performance advantages of DGPF increase with load, due to the greater diversity of scheduling combinations for users. Higher session initiation rate induces more traffic into the system. Therefore, the average throughput increases when more traffic flows are accepted, up to the transmission limit. Meanwhile, the received service rate for Class A users is consistent due to the QoS control in the scheduling algorithms.

B. Comparison to WGPF

Figure 5 shows the aggregate throughput performance for the DGPF scheduling algorithm for multi-class traffic. For comparison, the WGPF scheduling algorithm is also implemented in the same system. In the weighted algorithm, Class A users are granted higher priority via a weighted PF parameter.

We use the ratio of average throughputs for Class A and Class B as the control variable on the horizontal axis in Figure 5. Clearly, DGPF provides much greater throughput than WGPF. In WGPF, the weighted parameter gives priority to Class A users, but also compromises the average throughput, compared to the value achieved in PF algorithm. As a consequence, the overall throughput is degraded because low transmission efficiency slows down the packet arrivals due to the closed-loop transmission control (i.e., TCP) in the system.

C. Static versus Dynamic Control

To show the advantages of dynamic resource allocation in the scheduling algorithm, we compare two cases in the DGPF scheme. The results are displayed in Figure 6. The lower curve represents the results for an algorithm with a static L_A threshold, while the upper line represents an algorithm with a dynamic L_A threshold for resource allocation.

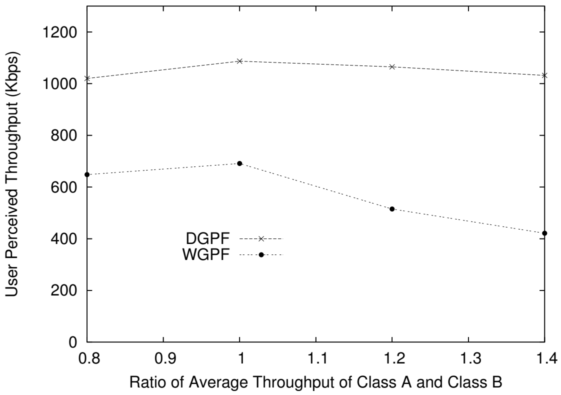
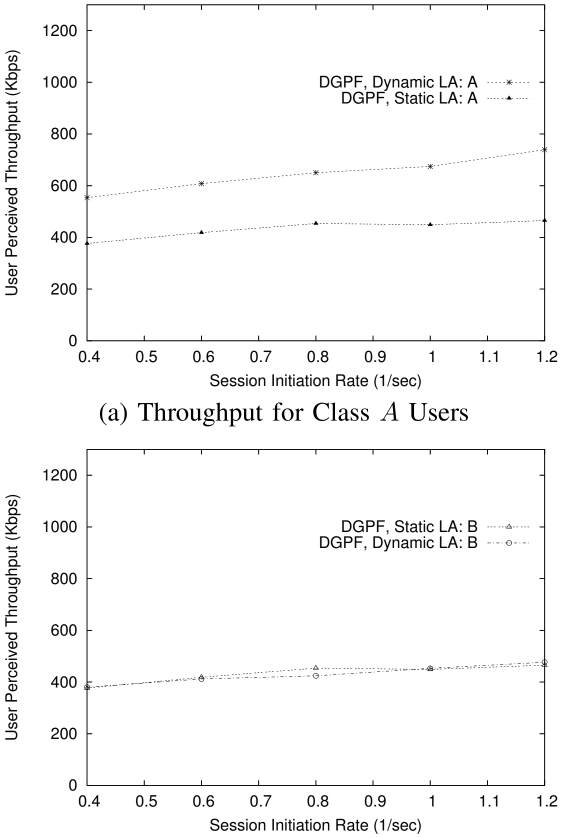


Fig. 5. Comparison of DGPF and WGPF

Figure 6(a) illustrates results for Class A users and Figure 6(b) is for Class B users. While Class B users are minimally affected, the dynamic threshold tends to improve the throughput for Class A users. Basically, it is better able to adapt to the rapid variations in channel quality and the bursty traffic arrival pattern at the user queues. Furthermore, it continues to achieve service differentiation between the user classes. In summary, the overall transmission efficiency is higher if the dynamic channel conditions are considered when doing QoS scheduling.



(a) Throughput for Class A Users

(b) Throughput for Class B Users

Fig. 6. Comparison of Static and Dynamic Resource Limits

D. TCP Dynamics

Our final set of simulation experiments considers the sensitivity of simulation results to TCP modeling details.

We consider small-scale simulation experiments with a single HTTP user, and focus on both queueing delay and throughput under several different TCP scenarios: persistent versus non-persistent connections, unconstrained versus receiver-limited window evolution, and per-packet versus per-connection Round-Trip-Time (RTT) modeling.

In general, these simulation experiments demonstrate the substantial throughput advantages of persistent connections, but with much higher mean and variance in the packet queueing delay. Per-packet RTT models also exacerbate this problem by adding greater variability to the end-to-end delay for object transfers. The main take-home message from these experiments is that the level of detail invested in the TCP model can dramatically affect the results obtained using an HSDPA simulator.

VII. Conclusion

This paper studies multi-user QoS scheduling for efficient transmission in a multi-rate transmission system. Dynamic Global Proportional Fairness (DGPF) scheduling is proposed for efficient scheduling with QoS guarantees. General traffic is considered in the algorithm, as well as scheduling for differentiated QoS.

We simulate the proposed algorithm in a HSDPA environment, and investigate the system performance compared to GPF scheduling. Our results show advantages for our DGPF algorithm, and provide insights on the impacts from stochastic traffic arrival process in the system, as well as the feedback-based control mechanisms of TCP.

We conclude that QoS issues and dynamic traffic effects have to be carefully considered in scheduling design. Overall transmission efficiency could be low if global performance impacts are not fully considered.

Acknowledgments

The authors would like to thank the MASCOTS 2009 reviewers for their encouraging feedback on the initial version of this paper. Financial support for this research was provided by NSERC (Natural Sciences and Engineering Research Council), iCORE (Informatics Circle of Research Excellence), and TELUS Mobility, via the NSERC/iCORE/TELUS Mobility Industrial Research Chair in *Wireless Internet Traffic Modeling* at the University of Calgary.

References

[1] 3GPP, "UTRA High Speed Downlink Packet Access (HSDPA): Overall Description", *TS 25.308 Release 5*.

[2] 3GPP2, "1xEVDO Evaluation Methodology v1.3", *3GPP2 C30-DOAH-20030818-004*, August 2003.

[3] 3GPP2, "CDMA2000 High Data Rate Packet Data Air Interface Specification", *3GPP2 C.S0024-A v2.0*, July 2005.

[4] R. Agarwal, V. Majjigi, R. Vannithamby, and J. Cioffi, "Efficient Scheduling for Heterogeneous Services in OFDMA Downlink", *Proceedings of IEEE Globecom*, November 2007.

[5] D. Amzallag, R. Bar-Yehuda, D. Raz, and G. Scalosub, "Cell Selection in 4G Cellular Networks" *Proceedings of IEEE INFOCOM*, April 2008.

[6] F. Angelis, F. Habib, I. Giambene, and G. Giannetti, "Scheduling for Differentiated Traffic Types in HSDPA Cellular Systems", *Proceedings of IEEE Globecom*, November 2005.

[7] G. Aniba and S. Aissa, "Adaptive Proportional Fairness for Packet Scheduling in HSDPA", *Proceedings of IEEE Globecom*, pp. 4033-4037, November 2004.

[8] C. Bettstetter, H. Hartenstein, and X. Perez-Costa, "Stochastic Properties of the Random Waypoint Mobility Model", *Wireless Networks*, Vol. 10 No. 5, pp. 555-567, September 2004.

[9] R. Cohen and L. Katzir, "Computational Analysis and Efficient Algorithms for Micro and Macro OFDMA", *Proceedings of IEEE INFOCOM*, April 2008.

[10] Q. Du and X. Zhang, "Resource Allocation for Downlink Statistical Multi-user QoS Provisionings in Cellular Wireless Networks", *Proceedings of IEEE INFOCOM*, April 2008.

[11] T. Girici, C. Zhu, J. Agre, and A. Ephremides, "Fair Scheduling in OFDMA-based Wireless Systems with QoS Constraints", *Proceedings of 30th International OFDM Workshop*, Hamburg, Germany, August 2007.

[12] J. Gomes, M. Yun, H. Choi, J. Kim, J. Sohn, and H. Choi, "Scheduling Algorithms for Policy Driven QoS Support in HSDPA Networks", *Proceedings of 26th IEEE Vehicular Technology Conference*, April 2007.

[13] A. Haider and R. Harris, "A Novel Proportional Fair Scheduling Algorithm for HSDPA in UMTS Networks", *Proceedings of 2nd International Conference on Wireless Broadband and Ultra Wideband Communications*, March 2007.

[14] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS, High Speed Radio Access for Mobile Communications*, Wiley, 2006.

[15] F. Kelly, A. Maulloo, and D. Tan, "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability", *Journal of the Operational Research Society*, Vol. 49, pp. 237-252, April 1998.

[16] H. Kim and Y. Han, "A Proportional Fair Scheduling for Multicarrier Transmission Systems", *IEEE Communication Letters*, Vol. 9, No. 3, March 2005.

[17] D. Kivanc, G. Li, and H. Liu, "Computationally Efficient Bandwidth Allocation and Power Control for OFDMA", *IEEE Transactions on Wireless Communications*, Vol. 2, pp. 1150-1158, November 2003.

[18] T. Kolding, K. Pedersen, J. Wigard, F. Frederiksen, and P. Mogensen, "High Speed Downlink Packet Access: WCDMA Evaluation", *Proceedings of IEEE Vehicular Technology Society News*, Vol. 50, No. 1, pp. 4-10, 2003.

[19] Y. Lu, C. Wang, C. Yin, and G. Yue, "Downlink Scheduling and Radio Resource Allocation in Adaptive OFDMA Wireless Communication Systems for User-Individual QoS", *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 12, March 2006.

[20] G. Song and Y. Li, "Utility-based Resource Allocation and Scheduling in OFDM-based Wireless Broadband Networks", *IEEE Communications*, pp. 127-134, December 2005.

[21] H. Sun and C. Williamson, "Downlink Performance for Mixed Web/VoIP Traffic in 1xEVDO Rev A Networks" *Proceedings of ICC*, Beijing, China, May 2008.

[22] J. Tsai, "State-dependent Proportional Fair Scheduling Algorithms for Wireless Forward Link Data Services", *Proceedings of IEEE INFOCOM*, April 2008.