Supporting Relative Delay Differentiation in CDMA Cellular Environments

Liu Wu and Ehab S. Elmallah
Department of Computing Science
University of Alberta
Edmonton, T6G 2H1, Canada
lwu@cs.ualberta.ca and ehab@cs.ualberta.ca

Abstract – This paper investigates the design and performance of a call admission control and a scheduling mechanism that aim at extending the relative delay differentiation per-hop behaviour of the Differentiated Services (DiffServ) model to 3G mobile users served by wide-band code-division multiple access (W-CDMA) air interface. Our study focuses on provisioning such service on the downlink of a system operating in the frequency-division duplex (FDD) mode. The design aims at maximizing the average effective throughput of each DiffServ class while satisfying prescribed relative delay constraints at congestion time. We compare the performance of a system with and without incorporating a simple mobility prediction in the call admission control phase. It is shown that both the predictive and the non-predictive schemes satisfy the relative delay differentiation constraints. The predictive scheme, however, results in decreased average delays for each class, and improves the effective throughput of lower priority classes.

I. Introduction

This paper considers extending the IETF's Differentiated Services (*DiffServ*) architecture for provisioning quality of service (QoS) to third generation (3G) wireless systems (see, for example, [1, 2, 3, 5, 9, 11] for information on DiffServ).

DiffServ classifies traffic into a small number of aggregate classes at the edge of each autonomous networking domain, and subsequently gives differential treatment to such classes inside the network at congestion time. Each user traffic is associated with a service level agreement (SLA) that specifies the expected traffic profile, as well as the expected forwarding service. Extending DiffServ to next generation wireless cellular systems is expected to provide a low cost means of running guaranteed-service applications on personal communication devices with traffic passing through the Internet.

In this paper, we focus on extending the *relative delay differentiation* [9] per-hop forwarding behaviour on the downlink of the wide-band code-division multiple access (W-CDMA) air interface operating in the frequency-division duplex (FDD) mode. (Other approaches for QoS provisioning in CDMA environments appear e.g. in [4, 6, 7, 8, 10].)

In the relative delay architecture, the traffic is classified into a fixed number of *delay classes*. At any given router, the ith delay class is associated with a *delay weight* Δ_i. In [9], the forwarding behaviour of a router is designed to ensure that the average delays perceived by packets in any two delay classes are in the inverse ratios of the corresponding delay weights. That is, if $\overline{d_i}(t)$ is the average delay incurred by class i packets over a time window of length t, then for any two classes i and j, $|\overline{d_i}(t)\Delta_i - \overline{d_j}(t)\Delta_j| \rightarrow 0$. In [9], the above relative forwarding behaviour is proposed to work in conjunction with a *per-flow end-to-end delay class adaptation* that dynamically adjusts the delay class of a flow in order to match the end-to-end delay requirements of the flow. The smaller the average delays encountered by a flow, the less frequent the need for the adaptation mechanism to switch the flow to a higher delay weight class. The adaptation aspect of the architecture, however, is not part of our study.

The remaining part of the paper is organized as follows. Section 2 gives an outline of the system model and the main parameters used in the simulation study. Section 3 describes two algorithms developed for our purpose: a relative delay differentiation scheduler and a call admission control algorithm. Section 4 presents some performance results.

II. System Model and Parameters

Throughout the paper we consider DiffServ traffic targeted to a central cell in a 19-cell configuration where the central cell is affected by the first- and second-tier interference caused by neighbouring base stations transmitting at maximum power. We assume a UMTS-like architecture where there is an Internet-gateway module that communicates with a designated DiffServ bandwidth broker and the base station of a target cell. We now describe the user mobility parameters, the traffic parameters, and the air interface parameters in the following sections.

A. *Mobility Parameters*

The mechanisms proposed in the next section for provisioning relative delay differentiation are evaluated under a

stressful scenario assuming a high user mobility model. For example, in our study we assume a random mobility model where a mobile user travels at an average speed of 10 m/sec, choosing one of eight possible directions every 3 seconds (as summarized in Table 1). To focus the study on the effect of the devised algorithms in such high mobility environment, we choose not to incorporate the effect of handoffs. The maximum possible number of active mobiles per cell determines the maximum number of concurrent flows the base station may deliver at any instant; the setting of this parameter is discussed below.

Parameter	Value	Unit
Cell radius	1000	m
Maximum number of mobiles per cell	40	
Average mobile speed	10	m/sec
Number of directions a mobile can follow	8	
Interval before a mobile picks a new direction	3	sec

Table 1. Cell and mobility parameters.

B. *Traffic Parameters*

In our simulation study, traffic generation is not intended to capture flows from any realistic application. Rather, the choice of the traffic parameters (together with the settings of the total base station transmission power available for the DiffServ traffic, and the maximum possible number of concurrent flows in the cell) are intended to ensure the generation of workload that is likely to cause no congestion at lower transmission rates (e.g. less than 64K bps), and a definite congestion at higher rates (e.g. greater than 128K bps). (Due to the limited space, however, we omit the argument that supports the above aspect.)

Each of the incoming traffic flows is assumed to arrive to the Internet-wireless gateway at a certain data rate and for a prescribed duration of time (as determined by an associated SLA).

To support bounded end-to-end delays necessary to run certain user applications smoothly, we associate with each packet a maximum acceptable delay limit: if a packet is delayed at the Internet-wireless gateway more than the associated maximum delay limit, that packet is considered useless for the end user application (and hence dropped from the system).

Furthermore, to support the policy of charging the mobile user on the basis of the useful received traffic (rather than on the sheer volume of received traffic), we associate with each flow an *effective* throughput delivery ratio ρ. If the ratio of packets delivered within the acceptable time limit to the total number of packets in the flow is below ρ (e.g., below 90% of total number of packets in a flow), then we count this as failure in delivering the entire flow. All packets in such a failed flow do not contribute to the *effective throughput* of the system.

Table 2 summarizes the main parameters used. The duration of each flow is uniformly distributed from 60 to 90 seconds (which allows the user to travel a significant distance within a 1000-meter cell radius while receiving the flow). For each mobile, the inter-arrival time between flows is exponentially distributed with a mean value of 30 seconds. If a packet within a flow is not delivered within a maximum acceptable delay limit of 6 seconds, the packet is dropped.

Parameter	Value	Unit
Flow duration	[60-90]	sec
Mean flow inter-arrival time	30	sec
Mean packet inter-arrival time	1	sec
Mean packet length	420	bytes
Maximum acceptable packet delay limit	6	sec
Effective throughput success ratio (ρ)	0.9	

Table 2. Traffic parameters.

C. *Air Interface Parameters*

We assume a standard W-CDMA parameters. Each mobile requires a target E_b/N ratio:

$$(E_b/N) = \frac{W}{R}\,\frac{P_{rx}}{\gamma\, I_{intra\_cell} \;+\; I_{inter\_cell} \;+\; \eta_0 W}$$

where W is the chip rate, R is the transmission bit rate of the coded data, P_{rx} is the power received from the serving base station, γ is the orthogonality factor, I_{intra_cell} is the interference power received by the mobile from the serving base station, I_{inter_cell} is the interference power received by the mobile from the neighbouring base stations, and η_0 is the white noise power spectral density.

Parameter	Value	Unit
Base station power budget	25	watts
Chipping rate	4.096	Mcps
Noise spectral density (η_0)	-174	dBm
Orthogonality factor (γ)	0.2	
Convolutional coding rate	1/3	
E_b/I_0 requirement	7	dB
Log-normal shadowing exponent (n)	4	
Log-normal shadowing standard deviation (σ)	5	dB

Table 3. Air interface parameters.

D. *Performance Measures*

For each delay class, we use the *average class delay*, and the *effective throughput* to assess performance. The average class delay is computed over all packets in all flows for some target delay class (the average also includes packets that are dropped because of exceeding the acceptable delay limit.)

As mentioned above, the effective throughput is defined with respect to threshold ratio ρ: we count a flow to be successfully delivered if the system manages to deliver at least ρ of the flow's packets prior to their expiry time; otherwise the system fails to deliver the flow. The effective throughput is then obtained by restricting our attention to the successfully delivered flows and ignoring the failed flows.

In our present context, we seek to develop a packet scheduling algorithm and a call admission control algorithm that maximize the average effective throughput of the served delay classes, while satisfying the relative delay constraints mentioned above at moderate congestion times, given a limited base station transmission power budget and the existence of user mobility.

III. Scheduling and Call Admission Control

A. A Proportional Delay Differentiation Scheduler

The packet scheduler is responsible for scheduling the transmission of the admitted flows to the base station for subsequent transmission to mobile end users.

Our scheduler modifies the algorithm devised in [9] for wireline routers in the following aspect. In the wireline case one can assign one queue to all flows belonging to each delay class. In each time slot, as many packets are transmitted from each queue (in a first-come first-served order) to satisfy the required relative delay constraints. Many packets may belong to the same flow.

In contrast, in a soft-capacity environment of a CDMA system, efficient use of resources may go against allocating a large proportion of the available base station bandwidth to some head of queue traffic if the traffic is destined to a few remote users. Hence, multiple concurrent transmissions to different users in each class should take place within any time slot.

The proposed scheduler discussed here takes the above aspect into consideration by keeping a queue for each active flow within each delay class. The algorithm uses two main parameters: τ_{sched} the length of a *scheduling cycle*, and n_{win} the number of packets used to approximate the average delay $\overline{d_i}$ of the ith delay class at the beginning of each scheduling cycle. As well, the algorithm assumes knowledge of the base station power available for transmitting the scheduled DiffServ traffic, and an estimation of the distance between the base station and each active mobile (to determine the resulting radio link attenuation factor).

At the beginning of each new scheduling cycle, the algorithm considers the delay classes in descending order of their *normalized delays* (the products $\Delta_i\overline{d_i}$), within each delay class the algorithm considers the head-of-queue packet of each active queue in descending order of their delays. Given the above ordering, the algorithm selects as many packets as possible subject to the estimated availability of the base station power. The selected packets are then forwarded to the base station of the target cell.

Fig. 1 shows the performance of the scheduler assuming no call admission control (i.e., the system admits all arriving flows) and the existence of three delay classes with delay weights $\Delta_1 = 4$, $\Delta_2 = 2$, and $\Delta_1 = 1$. Each flow in each class is to be transmitted to the end user at an average data rate R (the x-axis of Fig. 1). As the data rate R increases from a low rate of 16K bps to a higher rate of 192K bps the system goes through three phases: (a) for $R \leq 64K$ bps the system is well provisioned and most packets uniformly incur negligible delays (hence the observed delay ratios is almost one), (b) as R increases to 128K bps the system becomes moderately congested; few packets exceed the acceptable delay limit; here the system achieves the required delay ratio constraints, finally (c) as R exceeds 144K bps the system becomes heavily congested, most packets are dropped from low priority classes, and the per class average delay approaches the maximum acceptable delay limit value.

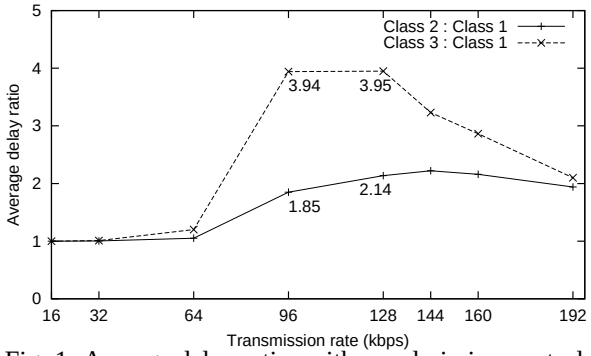

Fig. 1. Average delay ratios with no admission control

The objective of a carefully designed call admission control is to improve on the above performance by regulating the accepted flows in order to avoid operating the system in the heavily congested region, without sacrificing the effective throughput.

B. Call Admission Control

The call admission control (CAC) is responsible for regulating the admission of flows communicated through a designated DiffServ bandwidth broker, and forwarding the admitted packets to the scheduler mentioned above.

An important aspect of this study is the consideration of flows that have time duration long enough for the target end users to travel a considerable distance from the base station while receiving the flow. It is possible that, during such a time period, the users change their locations, and thereby require a total transmission power that exceeds the available base station power. If such power shortage occurs frequently many packets are likely to miss their delay expiry time, which will cause a loss in the system's effective throughput. To mini-

mize the risk of admitting flows that are likely to cause the base station to have a power shortage at some instants in the future, we adopt a simple mechanism that aims at predicting the base station power requirements as mobile users move near or away from the base station.

Now, suppose we would like to assess whether or not the base station will suffer from a power shortage during an interval t_{pred} in the future. We sample the space of outcomes in the following way: we conduct n_{trial} trials (e.g., $n_{trial} = 3$). Each trial involves a number of checkpoints (the checkpoints are equally spaced, and separated by a time interval $t_{interval}$ of prescribed length; thus, $\lfloor t_{pred}/t_{interval} \rfloor$ checkpoints are examined during a prediction interval of length t_{pred}).

At each checkpoint, we simulate the random movements of the users, and check whether there exists a feasible assignment of the base station power to each user. If there is no feasible power assignment at some checkpoint, then the corresponding trial fails (and there is no need to evaluate any remaining checkpoints in the trial). On the other hand, if a feasible power assignment exists for all $\lfloor t_{pred}/t_{interval} \rfloor$ checkpoints in a trial, then the trial succeeds. We consider each trial to be one sample, and design our predictive call admission control to accept a flow if the ratio between the number of successful trials (denoted as $n_{success}$) and the total number of trials n_{trial} exceeds a certain threshold, denoted as $p_{success}$.

We now illustrate the operation using some numerical values. Let us assume that the flow under test has a total length of $t_{flow} = 60$ seconds. Moreover, let us assume that the algorithm is set to sample the system for a prediction interval equals to 25% of t_{flow} (i.e., $t_{pred} = 0.25 * t_{flow} = 15$ seconds). If the algorithm performs checkpointing every $t_{interval} = 0.4$ seconds, then the algorithm considers approximately 37 checkpoints in each trial. If $p_{success} = 2/3$ and at least 2 out of the 3 trials succeed, the CAC algorithm accepts the new flow.

IV. Overview of Results

A. Average Class Delay

Fig. 2-a illustrates the average class delay obtained without incorporating call admission control. (Note: the ratios presented in Fig. 1 are based on the results of in Fig. 2-a.) As can be seen, when the data transmission rate for each active flow exceeds 144K bps the performance of the lower priority classes (2 and 3) deteriorate and approach the preset maximum acceptable limit.

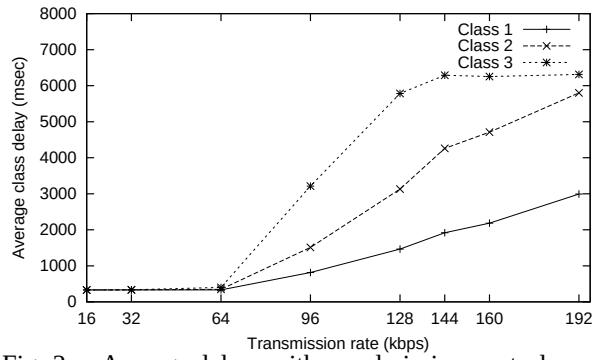

Fig. 2-a. Average delays with no admission control.

Figures 2-b and 2-c illustrate the impact of regulating the acceptance of the incoming traffic using a non-predictive CAC, and a predictive CAC respectively. Here, the predictive scheme uses a prediction interval that equals 10% of the flow duration time defined in the associated SLA. We recall from Section I that the smaller the average delays encountered by a flow, the less frequent the need for the end-to-end flow adaptation mechanism to switch the flow to a higher delay weight class.

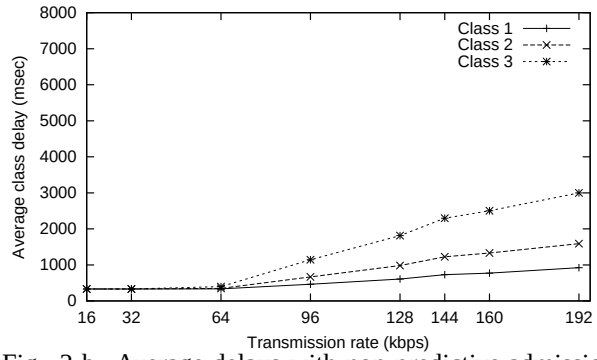

Fig. 2-b. Average delays with non-predictive admission control.

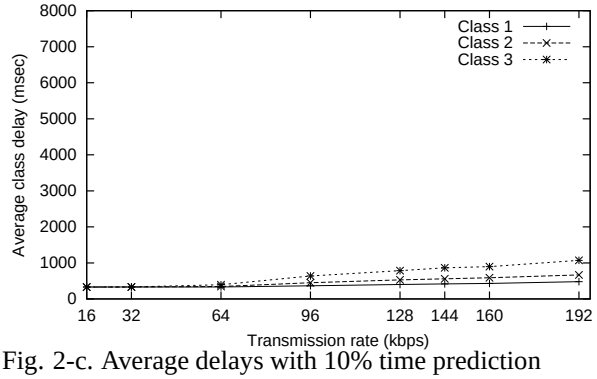

Fig. 2-c. Average delays with 10% time prediction

B. Effective Throughput

Similar to the average delay results, we present in Fig. 3-a the effective throughput of each delay class obtained without

admission control. As can be seen, the highest priority class (class 1) maintains uniform throughput level at the expense of the lower priority classes.

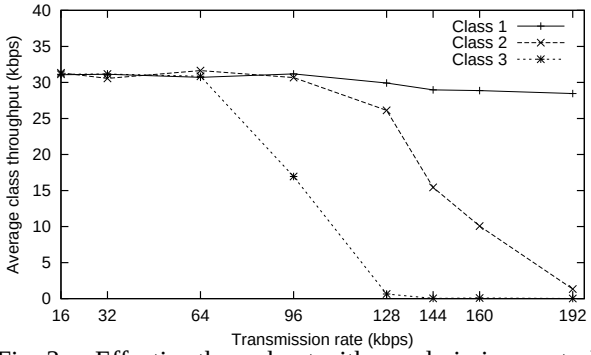

Fig. 3-a. Effective throughput with no admission control.

In contrast, using non-predictive admission control (Fig. 3-b) narrows the gap between the throughput of the different classes. The use of a predictive CAC (Fig. 3-c) results in further improvements in this regard.

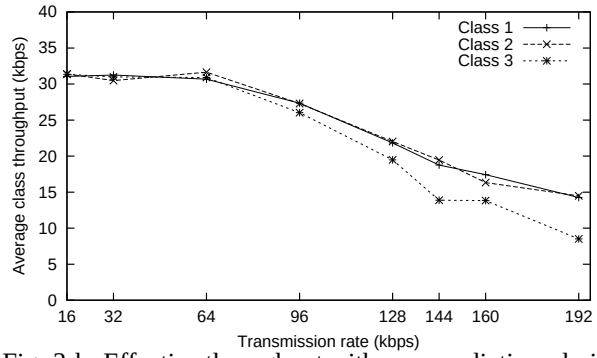

Fig. 3-b. Effective throughput with non-predictive admission control.

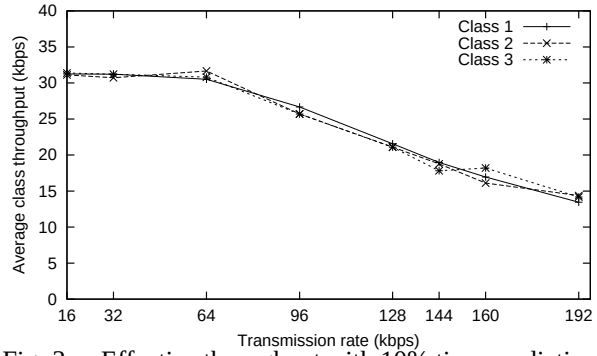

Fig. 3-c. Effective throughput with 10% time prediction.

7. Concluding Remarks

In this paper we develop and investigate the performance of a relative delay differentiation scheduler, and a predictive call admission control to provision the relative delay differentiation per-hop behavior of the DiffServ architecture on the downlink of a W-CDMA environment. Our simulation model uses evenly distributed traffic among three different delay classes. The results indicate that the predictive admission control improves the average delay of each class, and the effective throughput of low priority classes. Similar performance improvements have also been reported in [4] on the use of mobility prediction in provisioning the *assured forwarding* per-hop behaviour in W-CDMA environment. We are currently investigating the development of complementary mechanisms to extend the relative delay differentiation model to the uplink of a CDMA cellular wireless environment.

Acknowledgment

This research is supported by NSERC Canada and the Canadian Institute for Telecommunications Research (CITR). Part of this work has been done while the second author was visiting the Department of Computer Engineering at Kuwait University.

References

[1] S. Blake, D. Blak, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. IETF RFC 2475, December 1998.

[2] C. Dovrolis and P. Ramanathan. Proportional differentiated services, part ii: Loss rate differentiation and packet dropping. In *IWQoS*, June 2000.

[3] C. Dovrolis, D. Stiliadis, and P. Ramanathan. Proportional differentiated services: Delay differentiation and packet scheduling. In *SIGCOMM*, September 1999.

[4] E. Elmallah and H. Hassanein. A power-aware admission control scheme for supporting the assured forwarding model in cdma cellular networks. In *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks (LCN), Tampa, Florida*, 2003.

[5] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. Assured forwarding PHB group. IETF RFC 2597, June 1999.

[6] L. Jorguseski, J. Farserotu, and R. Prasad. Radio resource allocation in third-generation mobile communication systems. *IEEE Communications Magazine*, pages 117–123, 2001.

[7] N. S. Joshi, S. R. Kadaba, S. Patel, and G. S. Sundaram. Downlink scheduling in CDMA data networks. In *MobiCom 2000*, August 2000.

[8] M. Kazmi, P. Godlewski, and C. Cordier. Admission control strategy and scheduling algorithms for downlink packet transmission in WCDMA. In *52nd IEEE Vehicular Technology Conference*, September 2000.

[9] T. Nandagopal, N. Venkitaraman, R. Sivakumar, and V. Bharghavan. Delay differentiation and adaptation in core stateless networks. In *INFOCOM*, March 2000.

[10] D. Shen and C. Ji. Admission control of multimedia traffic for third generation CDMA network. In *INFOCOM 2000*, 2000.

[11] I. Stoica, S. Shenker, and H. Zhang. Core -stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks. In *SIGCOMM*, pages 118–130, 1998.