Achievable Quality-of-Service Guarantees in Wireless Communication Environments*

K.M. Tong, Mounir Hamdi, and Khaled Ben Letaief†
The Department of Computer Science
†The Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology
Kowloon, Hong Kong

Abstract In this paper, a multiple-cell wireless TDMA network is considered. In particular, we extend our previous work [1] for a single-cell wireless network in which multiple delay classes was mathematically analyzed and proved to be independent of the work-conserving scheduling algorithm used. Our extension includes the incorporation of a handoff model to provide an integration between call level and packet level quality-of-service (QoS) in a wireless TDMA environment. Throught extensive networking parameters, we verify the usefulness of our model in terms of its capability of determining the achievable QoS guarantees in a wirelss environment.

1 Introduction

It is expected that a significant portion of future networks' traffic will come from multimedia applications. Multimedia applications are different from traditional applications in that they require quality-of-service (QoS) guarantees in terms of delay, delay variation, and loss rate. In traditional applications, the system performance is largely measured in terms of the average overall throughput, average delay, and fairness and they are tolerant to network latencies. In contrast, real-time multimedia applications demand more stringent performance in temrms of QoS [?, 6]. In addition, their required QoS may vary from one application to another [4].

In this paper we focus on providing QoS on a wireless network. Typically, a wireless network consists of mobile devices, base stations, and the backbone network. A single base station can only cover a limited geographical area, or cell, and the mobile devices communicate with the base station using some radio frequencies in a shared manner. To enable communications between mobile devices of different cells, the base stations need to be connected, usually via a fast, wired backbone network so that the packets from the source mobile can be forwarded to the destination cell and transmitted to the receiving mobile.

Within a cell, all mobile hosts share the transmission medium using a certain Medium Access Control (MAC) scheme. Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) are two examples [3]. In this paper we focus on TDMA wireless networks. When considering the support of QoS on these wireless networks, we need to consider the provision of QoS within a cell and across multiple cells. These two aspects are dependent on each other [4].

In our previous paper [1] we have considered a single-cell system traffic of multiple delay classes. Then we mathematically analyzed and proved that it is independent of the scheduling algorithm used, for all work-conserving earliest-due-date (WC-EDD) scheduling algorithms. The dropping requirements of all individual applications are guaranteed using deadline-sensitive ordered-head-of-line (DSO-HoL) priority schemes. Verification of the model was shown through extensive simulations.

In this paper, we extend that research work to account for the more general case where we consider the provision of QoS across multiple cells. In this connection we emphasize on modeling the cells handoff schemes so as to account for the movement of mobile users between different cells while still being able to provide them with QoS.

The rest of the paper is organized as follows. In Section 2, we introduce the model for a single-cell wireless network. Section 3 extends this model to account for handoff. Section 4 presents experimental results of our model. Section 5 concludes paper.

2 Wireless Network Model

In TDMA wireless network, time is divided into fixed-sized frames. A frame can be divided further into slots, which can be fixed-size or variable-size. When an application tries to send a packet, it must ensure there are free slots available so as to avoid collisions with other packets. One way to ensure the availability of slots is by contention. Another way is by allocation in the base station (BS). Typically, there is a queue in every mobile host (MH) for holding ready-to-send packets. If the BS informs a MH about slot availability, the MH would select some packets from the queue for transmission. If real-time applications are to be supported, the queue would be a priority queue such that the MH selects the most critical packets first when slots are available. Figure 1 shows the general model under consideration.

Each MH is assumed to make requests to the BS at frame boundaries. The BS would schedule the available bandwidth to different MHs according to the degree of importance.

*This research work was supported in part by a Hong Kong Research Grant Council Grant.

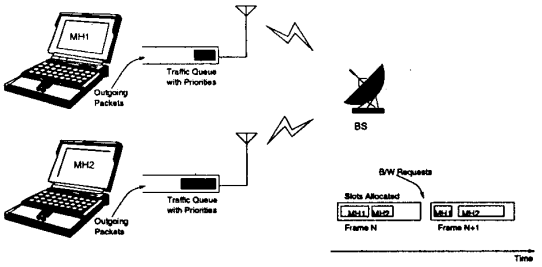

Figure 1: A general TDMA system

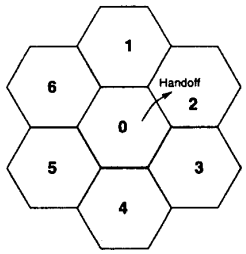

Figure 2: A handoff example from cell 0 to 2 with probability $h^0 \times H_{02}$.

In this paper, a generalization of the traffic model used in [3], which supports multiple delay requirements for different applications, is proposed. We assume all packets generated by a particular application have the same deadline, which is specified in terms of number of frames. An application may generate very time-sensitive packets which must be sent within the next frame or it must be dropped. Other applications, which can tolerate a longer delay, may still drop their packets if they are not serviced within, for instance, 4 frames. We define:

- A class N packet would be dropped if it cannot get serviced in the next N frames.
- Class N application would only generate class N packets.

N is basically the maximum tolerable delay parameter of an application which the network should guarantee. As a class N application generates class N packets, a class N packet may become a class $N-1$ packet if it cannot be serviced in the current frame, since the maximum delay tolerable of the packet is changed from N frames to $N-1$ frames. This kind of class $N-1$ packet is called the *residual packet*.

An application i is said to be of class c_i if it would only generate new packets that must be serviced within c_i frames. We define $\lambda_i(n)$ to be the number of new packets from application i entering the system at the beginning of frame n. We also define *residual packets* to be packets generated in previous frames which are neither expired (dropped) or serviced (transmitted). Let $r_i^c(n, f)$ be the number of residual packets at the beginning of frame n from application i that must be serviced within the next c frames, where f is the scheduling algorithm running at the base station.

The new arrival of all class c application form the system arrival of class c, defined as:

$$N_c(n) = \sum_{i,st.c_i=c} \lambda_i(n) \quad (1)$$

The total residual packets of class c in the system is:

$$R_c^f(n) = \sum_{i,st.c_i=c} r_i^c(n, f) \quad (2)$$

The total number of class c packets in the system is defined as:

$$\Lambda_c^f(n) = N_c(n) + R_c^f(n) \quad (3)$$

3 Handoff Consideration

In our previous paper [1], the deterministic bound on MAC level QoS for a single cell is considered. As an extension, we introduce a multiple-cell model in this paper. Hence, in addition to packet level QoS, we consider call level QoS. However, it is difficult to analyze the statistical call level QoS with deterministic MAC level QoS. A sophisticated bandwidth reservation scheme would be too complicated to be analyzed mathematically together with MAC level parameters. In this paper, a handoff model is first introduced, followed by the descriptions of some simple bandwidth reservation schemes and the performance analysis of a selected scheme.

3.1 Handoff Model

The movement of a person depends heavily on the current cell he or she is located. For instance, the handoff patterns would be significantly different when the person is on a highway compared to the one in a shopping mall. On the highway, it is expected that the handoff pattern of any given connection would be along the highway, unidirectionally. Whereas, in the shopping mall, a random pattern can be expected. Hence, the system parameters are defined based on this observation.

3.1.1 System Parameters

The following defines the parameters related to a cell:

- h^c defines the probability of handoff for a connection in cell c for each frame.
- H_{ij} defines the handoff preference such that, when handoff occurs, the probability of a connection going from cell i to cell j.

In Figure 2, for example, a handoff event at cell 0 occurs with probability h^0 for every TDMA frame. This parameter is analogous to the average walking speed of the mobiles at cell 0 since

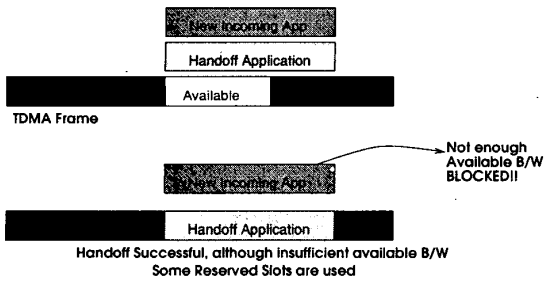

Figure 3: Example of the usage Bandwidth Reservation.

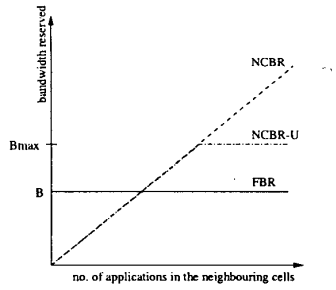

Figure 4: Comparison of the three reservation schemes.

when they are traveling fast, the handoff probability h^0 would be relatively higher, and vice versa. The parameter H_{02} shows the preference of the mobiles in the cell 0 that would handoff into cell 2, which is analogous to the directional component of the mobiles. Therefore, the probability of a handoff to happen from cell 0 to cell 2 is $h^0 \times H_{02}$.

Obviously, the handoff source shown in Figure 2 would increase the dropping rate in cell 2 as well as decrease the dropping rate in cell 0. Once the dropping rate in cell 2 is over the required limit, it means that the handoff call greatly affects the target cell such that the required dropping rate cannot be scheduled. Such a handoff call cannot be accepted and it should be dropped.

3.2 Bandwidth Reservation

Bandwidth reservation is key to lowering the handoff dropping probability. A good bandwidth reservation scheme could also make the network utilization high. However, a sophisticated reservation scheme would make the analysis extremely complicated. In this section, some general resource reservation schemes would be discussed which are applicable for our system model. It should be noted that handoff application is similar to a new incoming application, except it has a higher priority in the sense that the handoff application has a higher chance of success than the new incoming application under the same condition, as illustrated in Figure 3. Therefore, when a handoff occurs, it is treated in the same way as a new incoming application. The reserved bandwidth is used only by the handoff applications when there is insufficient bandwidth during the handoff.

3.2.1 Fixed Bandwidth Reservation (FBR)

For an FBR scheme, some fixed number of slots are being reserved exclusively for handoff application known as *guard channels*. If a frame has T slots, and there are B guard channels, the resource actually available would be reduced to $T - B$. When a new incoming application is requesting resource from the cell, admission rule applies with a reduced number of slots per frame $T' = T - B$.

When N applications handoff to the cell at the same time where the available slots cannot accommodate all of them, some need to be dropped. It is difficult to tell whether supporting 1 large application, while dropping all the other small applications, is better than the other way round. This problem exists in nearly all handoff-enabled systems.

3.2.2 Number-of-Connections-Based Reservation (NCBR)

There are a number of variations to this NCBR scheme. Basically, the bandwidth reserved is proportional to the total number of connections in all the neighboring cells. A proportional constant α is to be determined so that for every neighboring application, α slots are reserved.

Obviously when α is large, handoff dropping probability would be low. However, the bandwidth utilization of the cell would also be low. Tuning α to an optimal value is the key step for NCBR scheme. Normally an *upper bound* B_{max} is imposed on this scheme (NCBR-U) to limit the maximum reservation rate.

Under NCBR and NCBR-U, bandwidth is reserved *when necessary* and the number of slots reserved is changing over time, denoted by $B(n)$, according to the status of the neighboring cells. Admission rule is more complicated than FBR. The available resource of each cell is dynamically changing which is equal to $T' = T - B(n)$. An accurate estimation of $B(n)$ is very important in the admission rule.

Figure 4 illustrates the above-mentioned reservation schemes.

4 Analysis

In [1], the system dropping rate describes the packet dropping rate in a particular cell, and handoff is not considered. If it is to be extended, the system (cell) packet dropping rate, can be given in general by the following equation:

$$b_S = \mathrm{E}\left(\overline{\Lambda_1(n) + B(n)} - T \,\middle|\, \overline{\Lambda_1(n) + B(n)} > T\right) \times P\left(\overline{\Lambda_1(n) + B(n)}\right) \quad (4)$$

where $B(n)$ is the reserved bandwidth for handoff connections. Only handoff applications could use the reserved bandwidth. $\Lambda_1(n)$ is the total class 1 traffic in the cell. T is the total number

of slots per frame. This equation is very similar to equation given in [1], but with an additional term $B(n)$.

When a handoff application i, originally from cell k, requires to go into a cell c at frame n, cell c would try to support the application as a new application. If it cannot support it, bandwidth would be borrowed from the reserved bandwidth $B(n)$. Since the neighbor cell k has 1 application less, cell c should reconfigure according to the bandwidth reservation function.

The whole analysis lies on the prediction of $B(n)$ and $\Lambda_1(n)$. Table 1 shows the functions for $B(n)$ with corresponding reservation schemes. $\Phi_c(n)$ is the total number of applications in the neighboring cells of cell c, and α is a proportional constant.

$B(n)$	Reservation Scheme
K (Constant)	FBR
$\alpha\Phi_c(n)$	NCBR
$\min(\alpha\Phi_c(n), B_{max})$	NCBR-U

Table 1: Examples of Bandwidth Reservation Schemes with corresponding $B(n)$ functions

Let us define the following terms:

$\phi_c(n)$ is the number of applications in cell c at frame n.

$I_c(n)$, Number of Immigrants, is the number of handoff connections moving into the cell c at frame n.

$$I_c(n) = \sum_{\forall k \in \text{neighbour(c)}} H_{kc} \times h^k \times \phi_k(n-1) \qquad (5)$$

$E_c(n)$ is the number of Emigrants from cell c to another cell at frame n.

$$E_c(n) = h^c \times \phi_c(n-1) \qquad (6)$$

$N_c(n)$ is the number of new incoming connections in the cell c at frame n.

$D_c(n)$ is the number of departure from cell c at frame n.

From the definitions, the number of applications in cell c can be derived as the number of ongoing applications in the cell, plus net immigrants ($I_c(n) - E_c(n)$) and net arrivals ($N_c(n) - D_c(n)$). Equation (7) shows the relationship.

$$\phi_c(n) = \phi_c(n-1) + \overline{I_c(n) - E_c(n)} + \overline{N_c(n) - D_c(n)} \qquad (7)$$

It can be seen that from equation (5), (6) and (7), the number of applications in a cell at the current frame, $\phi_c(n)$, depends on all neighboring cells, $\phi_k(n-1)$ (k is a neighboring cell of c), and the current cell, $\phi_c(n-1)$, at the previous frame.

The expected number of applications in a cell, $\mathrm{E}(\phi_c)$, is calculated by iterations. Once $\mathrm{E}(\phi_c)$ is known, $\mathrm{E}(\Phi_c)$, the number of applications in the neighboring cells, is also known. Therefore, the expected bandwidth reservation, $\mathrm{E}(B(n))$, can be deduced.

4.1 Call Blocking Probability

If NCBR-U scheme is used, the bandwidth reservation is varying over time. Call admission, based on equation (4), depends on two variable factors: the current class 1 traffic in the current cell (a MAC level factor), and the current bandwidth reservation (a call level factor). Moreover, if heterogeneous traffic sources are used with different bandwidth requirements, the call blocking probability would not be a unique. This is because the call-level scheduler can always select many small bandwidth applications to admit rather than a large bandwidth application in order to lower the call blocking probability. This also adds another complexity to the analysis.

To simply the analysis, the following assumptions are assumed:

- Expected bandwidth reservation, $\mathrm{E}(B(n))$, is used for new call admission, rather than the actual process $B(n)$.
- All the applications are homogeneous with the same delay class and traffic distribution.

The first assumption makes the approximate analysis of NCBR-U the same as FBR, and the second assumption makes the value for call blocking probability unique without a specific call-level scheduler. This also implies that there would be a maximum number of applications, N_0, supported by a single cell. Hence, when a cell has less than N_0 applications, a new application originating in the cell would not be dropped. On the other hand, when there are N_0 applications in the cell, any new applications would be dropped. Therefore,

$$\text{Call Blocking Probability} = P(\phi_c(n) > N_0) \qquad (8)$$

4.2 Handoff Dropping Probability

Handoff dropping probability can be calculated in a similar fashion. Suppose the cell can support N_H simultaneous connections under no reservation. Handoff dropping occurs only if the cell has reached this limit. Therefore, handoff dropping can be estimated by:

$$\text{Handoff Dropping Probability} = P(\phi_c(n) > N_H) \qquad (9)$$

5 Numerical Examples

Assume each cell has a bandwidth of 1 Mbps. The homogeneous application under discussion is the voice source described in [1, 4]. The other system parameters can be found in table 2.

Assuming the system arrival rate and system departure rate are exponentially distributed, by tuning the system arrival rate, the average number of application in the reference cell, cell 0, can be obtained and is shown by Figure 5.

It is required that the bandwidth to be reserved in the cell equals to $\alpha\Phi_c(n)$, and the expected bandwidth reservation equals to $6\alpha \times \mathrm{E}(\phi_c(n))$ under the assumption that all the cells are statistically identical and symmetric. The expected bandwidth reservation is plotted in Figure 6. Based on the expected bandwidth

Parameter	Value
Number of Cells	2 rings (19 cells)
Max Voice Pkt Dropping	1 per frame
Handoff Prob (h_c)	0.3
Handoff Preference (H_{ij})	$\frac{1}{6}$
Proportional Constant (α)	1
Reservation Limit (B_{max})	10
Average System Departure Rate	10 apps per frame

Table 2: System parameters for handoff analysis

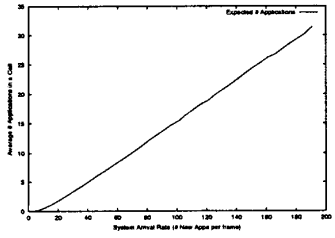

Figure 5: Average number of applications in a cell vs. System Arrival Rate.

reservation, the maximum number of applications that can be supported, N_0, can be calculated as shown in Figure 7 and the expected bandwidth is plotted again for reference.

From Figure 7, it is possible to estimate the call blocking probability and handoff dropping probability based on equation (8) and (9). The results are shown in Figure 8.

6 Conclusion

In this paper, we analyzed analyzing call level QoS together with packet level QoS presented in [1]. First, a handoff model is described, which is able to capture the speed and the direction of the mobile devices. Then, FBR, NCBR and NCBR-U reservation schemes are discussed and their corresponding conditions for the new call blocking and handoff dropping events are given. Performance analysis on the call blocking probability and handoff dropping probability is done under the NCBR-U schemes. Two assumptions are made to simply the handoff model for analysis. The first one is to replace the reservation process $B(n)$ by the corresponding expected value $\mathrm{E}(B(n))$. The other assumption assumes all applications to be homogeneous, which have the same bandwidth and delay requirement.

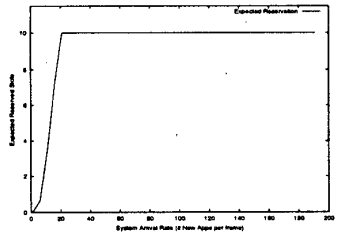

Figure 6: Expect bandwidth reservation vs. System Arrival Rate

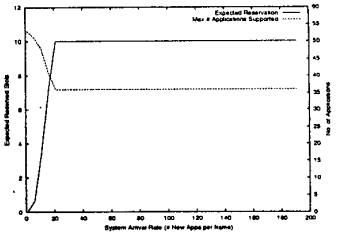

Figure 7: Maximum supported number of new applications vs. System Arrival Rate

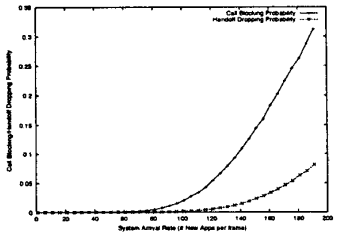

Figure 8: Call Blocking/Handoff Dropping Probabilities vs. System Arrival Rate

References

[1] K. M. Tong and M. Hamdi. Achievable QoS for Multiple Delay Classes in a Cellular TDMA Environment. *Proceedings of IEEE WCNC '2000,* Sept. 2000.

[2] J.G. Kim, I. Widjaja. Connection admission control for PRMA/DA wireless access protocol. *1997 IEEE Int. Perf., Comput. and Commun. Conf.,* 1997, pp.476-82.

[3] J.M. Capone, I. Stavrakakis. Delivering QoS Requirements to Traffic with Diverse Delay Tolerances in a TDMA Environment. *IEEE Trans. on Net.,* Feb 1999.

[4] K.M. Tong. Achievable QoS for Multiple Delay Classes in TDMA Cellular Environments. *MPhil Thesis,* Computer Science, Hong Kong University of Science and Technology, August 1999.

[5] S.S. Panwar, D. Towsley, J.K. Wolf. Optimal scheduling policies for a class of queues with customer deadlines. *Journal of the ACM,* Oct. 1988, pp.832-44.

[6] A. Acampora, M. Naghshineh. Control and QoS Provisioning in High-Speed cellular Networks. *IEEE Personal Commun.,* 1994.