threads, was discussed. In summary, communication servers which support QoS contracts are an important component of future QoS-aware services. This paper proposed new foundations for their design.

ACKNOWLEDGMENTS

The work reported in this paper was supported in part by the US National Science Foundation under grant nos. EIA-980620 and CCR-0093144 and US Defense Advanced Research Projects Agency grant N00014-01-1-0576.

REFERENCES

[1] T. Abdelzaher, "An Automated Profiling Subsystem for QoS-Aware Services," *Proc. Real-Time Technology and Applications Symp.,* June 2000.

[2] T. Abdelzaher and N. Bhatti, "Web Server QoS Management by Adaptive Content Delivery," *Proc. Int'l Workshop Quality of Service,* June 1999.

[3] T. Abdelzaher and K.G. Shin, "QoS Provisioning with *q* Contracts in Web and Multimedia Servers," *Proc. IEEE Real-Time Systems Symp.,* Dec. 1999.

[4] T.F. Abdelzaher, E.M. Atkins, and K.G. Shin, "QoS Negotiation in Real-Time Systems and Its Application to Automated Flight Control," *Proc. IEEE Real-Time Technology and Applications Symp.,* June 1997.

[5] T.F. Abdelzaher and K.G. Shin, "End-Host Architecture for QoS-Adaptive Communication," *Proc. IEEE Real-Time Technology and Applications Symp.,* June 1998.

[6] C. Aurrecoechea, A. Cambell, and L. Hauw, "A Survey of QoS Architectures," *Proc. Fourth IFIP Int'l Conf. Quality of Service,* Mar. 1996.

[7] G. Banga, P. Druschel, and J.C. Mogul, "Resource Containers: A New Facility for Resource Management in Server Systems," *Proc. Symp. Operating Systems Design and Implementation (OSDI),* 1999.

[8] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms,* chapter 25, pp. 527-531. The MIT Press, 1990.

[9] P. Druschel and G. Banga, "Lazy Receiver Processing (LRP): A Network Subsystem Architecture for Server Systems," *Proc. Symp. Operating Systems Design and Implementation (OSDI),* 1996.

[10] R. Gopalakrishnan and G. Parulkar, "Efficient User Space Protocol Implementations with QoS Guarantees Using Real-Time Upcalls," *IEEE/ACM Trans. Networking,* 1998.

[11] M. Jones, D. Rosu, and M.-C. Rosu, "CPU Reservations and Time Constraints: Efficient, Predictable Scheduling of Independent Activities," *Proc. 16th ACM Symp. Operating Systems Principles,* Oct. 1997.

[12] C. Lee, R. Rajkumar, and C. Mercer, "Experiences with Processor Reservation and Dynamic QoS in Real-Time Mach," *Proc. Multimedia,* Mar. 1996.

[13] I. Leslie, D. McAuley, R. Black, T. Roscoe, P. Barham, D. Evers, R. Fairbairns, and E. Hyden, "The Design and Implementation of an Operating System to Support Distributed Multimedia Applications," *JSAC,* June 1997.

[14] C. Mercer, S. Savage, and H. Tokuda, "Processor Capacity Reserves: Operating System Support for Multimedia Applications," *Proc. IEEE Int'l Conf. Multimedia Computing and Systems,* May 1994.

Thinning Schemes for Call Admission Control in Wireless Networks

Yuguang Fang, *Senior Member, IEEE*

Abstract—In this paper, we present new call admission control schemes, *the thinning schemes*, which smoothly reduce the traffic admission rates. Performance analysis is carried out and new analytical results are obtained. It demonstrates that the thinning schemes can be used to derive many known call admission control schemes.

Index Terms—Call admission control, Resource allocation, Wireless networks, Multimedia, Blocking probability.

✦

1 INTRODUCTION

THE future telecommunications networks (such as the third generation wireless networks) target providing integrated services, such as the voice, data, and multimedia, via inexpensive low-powered mobile computing devices over the wireless infrastructures ([1], [2]). The demand for multimedia services over the air has been steadily increasing over the last few years, leading to the design consideration of wireless Internet. Depending on the QoS (Quality of Service) requirements for various service requests from mobile users, different priorities may be assigned to various call connections. For example, real-time services such as voice or streaming videos may be assigned higher priority over non-real-time services such as data; handoff call connections should be given higher priority over new call connections in order to reduce the call dropping probability; mission critical data should be handled with higher priorities than some real-time data such as voice; users who pay more for their services should be treated with higher priorities over those who pay less. In order to support such mixed service requests in these wireless networks with multiple traffic types, efficient resource provisioning is a major issue ([2], [3]). Call admission control (CAC) is such a provisioning strategy to limit the number of call connections into the networks in order to reduce the network congestion and call dropping probabilities.

Prioritized traffic systems consisting of new calls and handoff calls in wireless networks have been intensively investigated in the literature (see [4], [5] and references therein). An admitted call for a mobile user may have to be handed off to another cell into which the mobile user moves, hence the call may not be able to gain a channel in the new cell due to the limited resource in wireless networks, which will lead to the call dropping. Thus, new calls and handoff calls have to be treated differently in terms of resource allocation. Since users tend to be much more sensitive to call dropping than to call blocking, handoff calls are normally assigned higher priority over new calls. The guard channel scheme ([3]) has been proposed to handle such systems: A proportion of the channels assigned for a base station has been reserved for handoff calls. This guard channel scheme can be generalized to handle the multimedia networks with multiple classes of priority services. Li et al. ([4]) have studied the guard channel scheme for wireless networks with multiple traffic types, the *multiple thresholding scheme*, in which different thresholds are used for each traffic type,

• *The author is with the Department of Electrical and Computer Engineering, University of Florida, 435 Engineering Building, PO Box 116130, Gainesville, FL 32611. E-mail: fang@ece.ufl.edu.*

Manuscript received 20 Sept. 1999; revised 14 June 2001; accepted 14 May 2002.

For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number 110621.

0018-9340/03/$17.00 © 2003 IEEE Published by the IEEE Computer Society

performance analysis has been carried out in certain detail. Recently, Ramjee et al. ([5]) proposed a so-called *fractional guard channel scheme* for the call admission control for wireless networks with two priorities traffic (new calls and handoff calls).

In wireless networks, a service area is populated with base stations, each is equipped with a number of channels for serving the mobile users in its coverage (*the cell*). Calls originating from the cell (called *new calls*) and calls handed over to the cell (called *handoff calls*) will share the resources (the channels) of the base station in the cell for services. To maintain the quality of service (QoS) for the ongoing calls (handoff calls), various priority schemes according to traffic types may be applied. In this note, we generalize *the fractional guard channel scheme* to handle multiple prioritized traffic streams in such wireless multimedia networks. We intend to use the priority levels to handle multimedia calls: A multimedia call connection may be granted higher priority over the voice and data services, a voice call may gain higher priority over a data call, and so on. In this way, higher prioritized call traffic will be given resources (channels) with higher probability in the proposed thinning schemes, hence the desired QoS can be met. Two variations have been studied and call blocking performance analysis has been carried out. It has been demonstrated that the new call admission control scheme, *the thinning schemes*, includes the guard channel scheme, fractional guard channel, and multiple thresholding scheme as special cases and can be used to obtain new CAC schemes for multimedia networks.

2 Thinning Schemes

Consider a wireless network which can support multiple types of services. To provide the desired QoS for each service, the network assigns channels according to the priority level. For example, at a base station (BS), if there are r traffic types of call arrivals (including new calls and handoff calls), each traffic type will be assigned one priority level. Thus, the BS can determine whether an arriving call connection is granted a channel or not based on the priority level.

The thinning scheme proposed in this paper is the one in which a call is admitted with certain probability based on the priority and the current traffic situation. The idea behind this scheme is to smoothly throttle the call arrival streams according to the priority as the network traffic is building up, thus, when the network is approaching congestion, the call streams with lower priorities become thinner. Due to the flexible choice of such call admission probabilities, these schemes can be made very general. In this section, we study two thinning schemes, the first one uses the information about the total number of busy channels, while the second scheme utilizes the numbers of channels occupied by the individual prioritized traffic streams.

We start with the study of the first scheme (Thinning Scheme I). The basic idea behind this scheme was first proposed by Ramjee et al. ([5]) for two call streams in wireless networks. We generalize this idea for multiple prioritized traffic in wireless multimedia networks in this paper. Assume that the wireless multimedia network has call requests of r priority levels, that each base station has C channels (or bandwidth units (BU) if the spectrum is not channelized as in the traditional cellular networks). Let α_{ij} ($i = 0, 1, \ldots C$ and $j = 1, 2, \ldots, r$) ($\alpha_{Cj} = 0$) denote the nonnegative numbers in the interval $[0, 1]$. The *Thinning Scheme* I (TS I) works as follows: When the number of busy channels at a BS is i, an arriving jth-stream call will be admitted with probability α_{ij} for $i = 0, 1, 2, \ldots, C$ and $j = 1, 2, \ldots, r$. All calls will be blocked when all channels at the BS are busy. The choice of these probabilities α_{ij} can be made according to the traffic situation (the number of busy channels): We choose α_{ij} small when i is large, which reflects the fact that, when a BS approaches a congestion state, some of the call arrivals are immediately rejected, even though there are still some channels available so that some higher priority calls can still be accepted later on. Fractional guard channel is a special case: We only have two priority levels (for new calls and handoff calls) and the call connections are for voice services.

To obtain tractable analytical results, we make the following assumptions throughout the paper: Call arrivals of each priority level form a Poisson process with arrival rate λ_j (for the jth priority call stream) and is independent of call arrival processes at other priority levels and the service times, all calls have channel holding times which are exponentially distributed with parameter μ (the channel holding time is defined as the time a call connection occupied in a cell).

The TS I can be characterized by one-dimensional Markov chain in which the state variable is the number of busy channels at the BS. It is observed that the transition rate from state i to state $i+1$ is $\sum_{j=1}^{r} \alpha_{ij}\lambda_j$, while the transition rate from state $i+1$ and i is $(i+1)\mu$. Let $\rho_j = \lambda_j/\mu$ for $j = 1, 2, \ldots, r$, let p_i denote the stationary probability at state i, we have

$$p_i = \frac{\prod_{k=0}^{i-1}\left(\sum_{j=1}^{r} \alpha_{kj}\rho_j\right)}{i!} p_0,$$

where

$$p_0 = \left[\sum_{i=0}^{C}\left(\frac{\prod_{k=0}^{i-1}\left(\sum_{j=1}^{r} \alpha_{kj}\rho_j\right)}{i!}\right)\right]^{-1}.$$

From this stationary distribution, we obtain the blocking probability for the jth priority call stream as follows:

$$P_j^b = \sum_{i=0}^{C}(1-\alpha_{ij})p_i, \quad \alpha_{Cj} = 0, \quad j = 1, 2, \ldots, r. \tag{1}$$

Obviously, when $\alpha_{0j} = \cdots = \alpha_{(m_j-1)j} = 1$ and

$$\alpha_{m_j j} = \cdots = \alpha_{Cj} = 0,$$

this scheme becomes the guard channel scheme (the cutoff priority scheme) in which m_j serves as the threshold for the jth priority calls: Whenever the number of busy channels is less than m_j, a new arriving jth priority call is accepted, otherwise rejected. When $j = 1$ corresponds to the new calls and $j = 2$ to handoff calls, if we choose $\alpha_{i2} = 1$, which implies that handoff calls are always accepted as long as there are channels available, the TS I reduces to the guard channel (cutoff priority) scheme commonly used in wireless cellular systems. We also observe that, when $\alpha_{0j} \geq \alpha_{1j} \geq \cdots \geq \alpha_{Cj}$, the jth priority call stream becomes thinner and thinner as the number of busy channels increases. Take the previous case for the wireless cellular systems with call stream of new calls and handoff calls, for example, assuming that $C = 4$, we choose $\alpha_{01} = \alpha_{11} = 1$, $\alpha_{21} = 0.5$, $\alpha_{31} = 0.3$, and $\alpha_{41} = 0$, then this CAC implies that, when there are no or one ongoing call in the cell, all new calls will be accepted, that, when there are two ongoing calls in the cell, a new call will be accepted with probability 0.5 and will be rejected with probability 0.5 even though there are still channels available, that, when there are three ongoing calls in the cell, a new call will be accepted with probability 0.3, a much lower probability than in the previous case, and that, when all channels are busy, a new call will be blocked. In terms of traffic load, the admitted new call stream is thinner and thinner as the number of ongoing calls becomes larger and larger. Thus, by appropriately choosing the parameters α_{ij} ($i = 0, 1, \ldots, C$, $j = 1, 2, \ldots, r$), we can throttle the call admissions accordingly.

A variation of the TS I is to admit the new arriving *j*th priority calls based on the number of the *j*th priority calls currently in service in the cell, we call it *the Thinning Scheme II (TS II)*. Let β_{ij} ($i = 0, 1, \ldots, C$ and $j = 1, 2, \ldots, r$) ($\beta_{Cj} = 0$) be nonnegative numbers in the interval $[0, 1]$. A newly arriving *j*th priority call is admitted with probability β_{ij} if there are i calls with *j*th priority currently in service in the cell and all calls will be blocked if all channels are busy. It is expected that the performance analysis has to be carried out using *r*-dimensional Markov chain theory. The state is now the vector $(m_1, m_2, \ldots, m_r)$, where m_j is the number of the *j*th priority calls currently in service in the cell. Using the standard argument as for queuing networks, we can easily obtain the stationary probability distribution:

$$p(m_1, m_2, \ldots, m_r) = \left[\prod_{j=1}^{r} \left(\frac{\prod_{i=0}^{m_j-1} (\beta_{ij}\rho_j)}{m_j!} \right) \right] \cdot p(0, 0, \ldots, 0),$$

$$m_1 + m_2 + \cdots + m_r \leq C,$$

where

$$p(0, 0, \ldots, 0) = \left[\sum_{m_1+\cdots+m_r \leq C} \prod_{j=1}^{r} \left(\frac{\prod_{i=0}^{m_j-1} (\beta_{ij}\rho_j)}{m_j!} \right) \right]^{-1}.$$

Thus, the blocking probability for the *j*th priority calls is given by

$$P_j^b = \sum_{i=0}^{C} (1 - \beta_{ij}) \left[\sum_{m_1+\cdots+m_{j-1}+m_{j+1}+\cdots+m_r \leq C-i} p(m_1, \ldots, m_{j-1}, i, m_{j+1}, \ldots, m_r) \right], \quad j = 1, 2, \ldots, r. \tag{2}$$

We observe that, when $\beta_{0j} = \cdots = \beta_{(m_j-1)j} = 1$ and $\beta_{m_jj} = \cdots = \beta_{Cj} = 0$, we obtain the bounding scheme (the multiple thresholding scheme): A *j*th priority call will be rejected if the threshold for *j*th type calls is reached.

As a final remark, we point out that the advantage of the thinning schemes is the parameterization of call admission control schemes so that optimization can be formulated. Such an optimization problem will be investigated in the future.

3 Conclusions

In this paper, we investigate call admission control strategies for the wireless networks. These schemes, the thinning schemes, can be shown to be general enough to include previously known schemes as special cases. By appropriately choosing the parameters in the schemes, they can smoothly throttle the call admissions based on the priority levels. This scheme should be useful for the ever increasing multimedia services support with various QoS requirements in the wireless environments.

Acknowledgments

This work was supported in part by a US National Science Foundation Faculty Early Career Development Award under grant ANI-0093241 and the US Office of Naval Research Young Investigator Award under grant N000140210464.

References

[1] Y. Fang, I. Chlamtac, and Y.B. Lin, "Channel Occupancy Times and Handoff Rate for Mobile Computing and PCS Networks," *IEEE Trans. Computers,* vol. 47, no. 6, pp. 679-692, June 1998.

[2] D. Grillo, R.A. Skoog, S. Chia, and K.K. Leung, "Teletraffic Engineering for Mobile Personal Communications in ITU-T Work: The Need to Match Practice and Theory," *IEEE Personal Comm.*, vol. 5, pp. 38-58, Dec. 1998.

[3] D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-Prioritized Handoff Procedures," *IEEE Trans. Vehicular Technology,* vol. 35, no. 3, pp. 77-92, 1986.

[4] B. Li, C. Lin, and S. Chanson, "Analysis of a Hybrid Cutoff Priority Scheme for Multiple Classes of Traffic in Multimedia Wireless Networks," *Wireless Networks (WINET),* vol. 4, no. 4, Aug. 1998.

[5] R. Ramjee, D. Towsley, and R. Nagarajan, "On Optimal Call Admission Control in Cellular Networks," *Wireless Networks,* vol. 3, pp. 29-41, 1997.