

Optimizing Medium Access Control For Rapid Handoffs In Pseudocellular Networks

R. Mudumbai

Department of Electrical and
Computer Engineering
University of California
Santa Barbara, CA 93106
Email: raghu@ece.ucsb.edu

G. Barriac

Department of Electrical and
Computer Engineering
University of California
Santa Barbara, CA 93106
Email: barriac@engineering.ucsb.edu

U. Madhow

Department of Electrical and
Computer Engineering
University of California
Santa Barbara, CA 93106
Email: madhow@ece.ucsb.edu

Abstract—The steadily decreasing cost of Wireless Local Area Network (WLAN) technology motivates the concept of *pseudocellular networks* that support real-time traffic and seamless mobility with WLAN-type infrastructure, in addition to the standard low-mobility data applications of WLANs. A key challenge in such networks is the support of highly mobile users with real-time traffic, because of the high handoff rate resulting from small cell sizes. In this paper, we show that timely mobile-centric handoffs can be achieved using optimized ALOHA-like reservation schemes which, for a given call drop probability, require significantly fewer reservation resources than conventional methods. For Poisson handoff traffic, we employ dynamic programming to derive an optimal stationary policy. We then use these results as a building block for obtaining adaptive reservation schemes (dynamically varying the number of minislots per frame) for bursty handoff traffic.

I. INTRODUCTION

With the availability of 802.11-based Wireless Local Area Network (WLAN) technology at mass market prices, the possibility of offering real-time services such as voice, in addition to the data services supported currently, becomes an economically attractive proposition. In this paper, we consider *pseudocellular* networks offering seamless mobility for voice, data and multimedia services across an entire university or industry campus, using WLAN technology alone. The campus would be covered by a backbone of standard WLAN Access Points (APs), each having a range of the order of about 100 meters (see Figure 1). Mobile nodes receive service from the AP nearest to them just like in a standard cellular system.

The focus of this paper is on supporting highly mobile users with real-time traffic (e.g., mobile telephony at vehicular speeds). The difficulty in this task arises because of the frequent handoffs caused by the small cell sizes. We propose to solve this problem by absorbing the task of handoffs into Medium Access Control (MAC), with a user entering a new cell simply asking the new AP for a reservation. Our contribution is to devise low-complexity ALOHA-like reservation schemes that are optimized for minimizing the

probability of a dropped call. We show that such mobile-centric handoff schemes can indeed support voice connections at vehicular speeds with reasonable overhead.

A. Related Work

Other researchers ([1], [2]) have studied real-time applications in 802.11 networks and identified the CSMA based MAC protocol ([3], [4]) as the significant source of latency and overhead. We propose to get around this limitation by employing a centralized TDMA based scheduling for uplink data using the Point Coordination Function (PCF), or equivalent software emulations, e.g. [5].

The idea of using contention based algorithms to make reservation requests for a TDMA based data channel is also well-known. In [6], Section 4.5, the authors show that under such a scheme, arbitrarily high throughputs can be achieved independent of the contention algorithm used, provided there are no delay constraints.

A reservation protocol similar to ours, has been standardized for upstream data transmission over cable TV infrastructure [7], [8]. The ALOHA system ([9]) itself has been studied extensively in the literature. In [10] and [11], the authors propose a multi-access channel model very similar to the one analyzed in this paper, where each contending node transmits multiple copies of its packet to provide a simple form of redundancy over the contention channel. In [12], the authors propose a splitting algorithm with QoS support for multiservice WLANs. The main difference of our work from previous work is that we derive ALOHA based multi-access schemes that minimize the *probability of missing delay deadlines*. This is different from the usual objective of maximizing channel utilization or throughput and leads naturally to adaptive strategies of bandwidth allocation for the reservation channel.

In [13], the handoff problem in pseudocellular networks is addressed by employing joint PHY/MAC optimization to obtain a high capacity reservation channel. This, however, requires more PHY capabilities than provided for in the 802.11b standard. By contrast, the MAC changes we propose can be achieved by firmware and software upgrades.

The paper is organized as follows: Section II describes the conceptual model of the reservation system and relates

This work was supported by the Office of Naval Research under grant N00014-03-1-0090, and by the National Science Foundation under grants ANI 0220118 and EIA 0080134

it to a pseudocellular network. Section III outlines a dynamic programming procedure to derive optimal reservation policies for the case of a) a slow Poisson arrival process and b) a bursty arrival process. Section IV discusses possible directions of future work and concludes the paper.



Fig. 1. Pseudocellular campus area network

II. SYSTEM MODEL

We consider the contention problem of mobile nodes trying to gain access to the 802.11b AP polling function by sending a reservation request. In general, each node is considered as belonging to one of a set of possible priority classes. Any set of contending users that can be separately addressed by the AP can be considered as a separate priority class. Thus we can have roaming users in a separate priority class from new users, and similarly users can be classified into priority classes depending on how close they are to their real-time deadline constraint.

A. A pseudocellular network

As an extreme case of reservation traffic, we consider the example of fast moving nodes maintaining long duration sessions of voice over IP traffic. If we have APs spaced at 100m apart, we expect a user moving at automobile speeds to roam into a new cell every 10 seconds or so. Assuming the voIP conversation ([14]) generates one 200 byte packet every 50 ms, the mobile is able to transmit 200 packets between handoffs. If handoffs are completed within a delay constraint of about 200ms, then with a small amount of buffering (4 voIP packets) we can assure that the connection will not drop any packets. Assuming that the AP allocates a reservation frame every 50 ms, each connection has a deadline of 4 reservation frames in which to complete the handoff.

If we utilize an 11 Mbps 802.11b link at 20% utilization, we should be able to sustain 50 such conversations per cell simultaneously. As the 802.11b standard evolves to higher speeds, this scenario should also improve significantly. Therefore we consider handoff rates of ≤ 0.25 users per reservation frame. In the course of a typical voice conversation (of ≈ 200 seconds duration), there would be about 20 such handoffs. If we want to maintain 1% drop rates over 20 handoffs, we require per-handoff probability of failure to be of the order of 10^{-3} .

B. Assumptions

We outline the main assumptions of our reservation model:

- 1) The pseudocellular network consists of 802.11b WLANs operating in Infrastructure Mode ([3]). Each mobile node is served by an AP acting as a Base Station in the network. The APs are interconnected by a separate wired backbone network, which has high bandwidth and low latency compared to the 802.11 link.
- 2) The mobiles use a polling based TDMA scheme for the uplink i.e. each mobile waits for its AP to assign a time-slot to it before transmitting data.
- 3) The handoff process is mobile-centric; the mobiles are *frequency agile*, and are able to decide when to perform handoffs by measuring signal strengths of nearby APs transmissions.
- 4) A mobile initiates a new call or a handoff by sending a reservation request on zero or more of several random access reservation minislots (Figure 2) following a beacon frame transmitted by the AP. The mobiles use the random access transmission policy specified by the AP in its beacon frame.
- 5) The AP is able to estimate the size of an initial arrival burst and also the multiplicity of each collision, so that it knows the size of each priority class at all times. 802.11b APs do not have this capability, however we can use an indirect approach: a) AP uses worst case estimates on size of a burst arrival, and b) AP assumes that all collisions are two node collisions. We show using simulation results, that for the optimal transmission policies derived in this paper, such crude estimates give excellent results.

III. OPTIMAL RESERVATION SCHEMES

We model the reservation process as a discrete-time system; there is a *reservation frame* every T seconds. For simplicity, we consider priority scheduling based only on each node's deadline, assumed to be A reservation frames for all nodes. In Section III-A, P_{fail} , the probability a node fails to make a reservation in A attempts, is the function to be minimized. By contrast, in Section III-B, we minimize the total reservation bandwidth (in number of minislots) required to assure a certain required P_{fail} .

We allow each *reservation frame* to consist of multiple *minislots*. A *minislot* is a time-slot for transmitting a MAC frame containing a reservation request. Hence, a *reservation frame* consists of a succession of such time-slots (see Figure 2), each separated by a small Inter-Frame Spacing which is required for physical layer synchronization in the 802.11 system ([4], [3]). Note that reservation frames can be dynamically extended to support higher handoff traffic by adding more minislots because the uplink scheduling is centrally coordinated by the AP.

A. Poisson arrival process

Given a stationary, Poisson arrival process of rate λ , roaming nodes with a deadline of A frames and a constant allocation of K minislots (parallel reservation channels) per

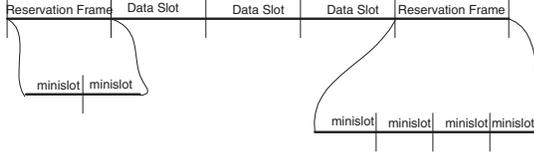


Fig. 2. Reservation frame model: minislots structure

frame, we derive the optimum transmission strategy for each node as a function of the delay deadline for that node.

We consider all nodes who have failed to access the channel in $i - 1$ consecutive reservation frames as belonging to class i , and denote the number of class i nodes in the system by N_i . There are exactly A classes of nodes since users who have failed A times must leave the system. As discussed previously, the AP is assumed to have an accurate estimate of the state vector $\bar{N} \equiv [N_1, N_2, \dots, N_A]$.

A *transmission vector* is defined as a binary vector of size K : $\bar{t} \equiv [t_1, t_2, \dots, t_K]$, where $t_j = 1$ means transmit in minislots j and $t_j = 0$ means idle in minislots j . There are 2^K possible transmission vectors corresponding to choosing a subset of K minislots to make a reservation attempt. We enumerate these vectors as $\bar{t}_1 \dots \bar{t}_{2^K}$.

A transmission policy, \bar{p}_i , is defined as a probability distribution over all possible transmission vectors for a node of class i , i.e. $\bar{p}_i \equiv [p_i(\bar{t}_1) p_i(\bar{t}_2) \dots p_i(\bar{t}_{2^K})]$ and $\bar{p} \equiv [\bar{p}_1 \bar{p}_2 \dots \bar{p}_A]$

We model this system as a *stationary infinite horizon Markov decision process* ([15], Chap. 7) with state \bar{N} , and consider the problem of minimizing the number of nodes failing to make a reservation in A attempts. A reservation is said to have succeeded if the node's transmission vector had a transmission without collision in at least one minislots.

We simplify the problem by truncating the state space so that no class has more than N_{max} nodes. We can then enumerate the state-vectors as $\bar{N}_1 \dots \bar{N}_L$ where $L = (N_{max} + 1)^A$.¹

Let \bar{p}_i^* be the optimal transmission policy and v_l be the cost of being in state \bar{N}_l , and let $\bar{v} \equiv [v_1 v_2 \dots v_L]$

$$\bar{p}_l^* \equiv \arg \min_{\bar{p}} C(\bar{N}_l, \bar{p}, \bar{v}) \quad (1)$$

where $C(\bar{N}_l, \bar{p}, \bar{v})$ is the cost (in expected number of failed users) of using policy \bar{p} when in state \bar{N}_l and the state cost vector is \bar{v} . More precisely, $C(\bar{N}_l, \bar{p}, \bar{v}) = N_{fail}(\bar{N}_l, \bar{p}) + \gamma(\bar{N}_l, \bar{p}, \bar{v})$, where $N_{fail}(\bar{N}_l, \bar{p})$ is the expected number of class A users that fail in the next reservation frame given the state is \bar{N}_l and the policy is \bar{p} , and $\gamma(\bar{N}_l, \bar{p}, \bar{v})$ is the cost associated with moving to the next state (averaged over all possible "next states").

We use an iterative procedure to compute the optimal \bar{p}^* . Letting k be the iteration index, we can write

$$\bar{p}_l^*(k) = \arg \min_{\bar{p}} C(\bar{N}_l, \bar{p}, \bar{v}(k)) \quad l = 1..L \quad (2)$$

$$\bar{v}(k+1) = \bar{C}^*(k) \quad (3)$$

¹We find that $N_{max} = 2$ is a good approximation for slow Poisson arrivals ($\lambda \leq 0.25$) considered in this section

where the arbitrarily chosen "initial cost vector" is $\bar{v}(1) = \bar{0}$, $\bar{C}^*(k) \equiv [C_1^*(k), \dots, C_L^*(k)]$ and $C_l^*(k) \equiv C(\bar{N}_l, \bar{p}_l^*(k), \bar{v}(k))$

It can be shown that this *value iteration* algorithm gives a cost function $\bar{C}_l^*(k)$ satisfying, $\frac{\bar{C}_l^*(k)}{k} \rightarrow \langle N_{fail} \rangle$ as $k \rightarrow \infty$ ([15], page 318). $\langle N_{fail} \rangle$ is the expected number of failed users per reservation frames averaged over all time, its value is independent of the initial state vector and can be estimated from the iterated values of the cost function, $\bar{C}_l^*(k)$ as $\frac{\bar{C}_l^*(k)}{k}$. The overall probability of failure is $P_{fail} = \frac{\langle N_{fail} \rangle}{\lambda}$.

Each iteration of the dynamic program involves solving an optimization problem with $A * (2^K - 1)$ independent variables for each value of the state-vector \bar{N} . We now state the following theorem which reduces the complexity of this problem significantly (the proof is omitted for lack of space).

Theorem 1: To minimize P_{fail} , it is optimal to have exactly one class of users contending in each minislots. \square

According to Theorem 1 it is always optimal to partition the available number of minislots into smaller number of minislots separately allocated to each class of contending nodes.

We thus reduce our problem to:

- 1) Find the optimal partition $[K_1, K_2, \dots, K_A]$ such that $\sum_j K_j = K$
- 2) Given N_i, K_i , for each class, find the optimal transmission policy \bar{p}_i (Equation 1), where \bar{t} is now a binary vector of size K_i .

In [11], the authors state and prove a related result for the optimality of a *pure transmission policy*² for a single class of users given N_i and K_i . Using this result, in Step 2 above we only have to consider the transmission policies which involve a fixed number of transmissions, R , for different values of R until we find the value that gives the minimum.

Figure 3 shows the comparison of our optimized scheme against simple slotted ALOHA for different λ , using 2 minislots per frame and a deadline of 2 frames.

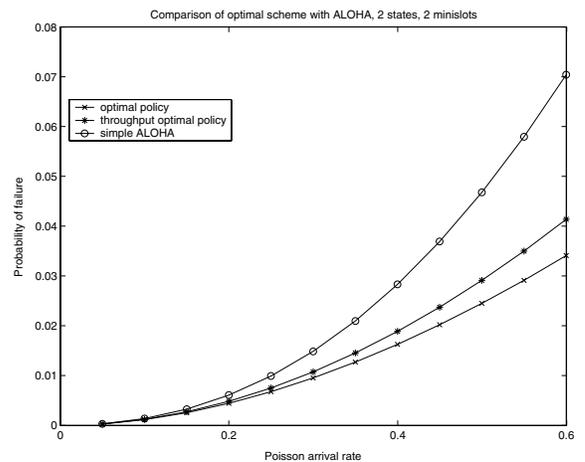


Fig. 3. Comparison of optimal policy with simple ALOHA

²A "pure" policy is one where each node's *transmission vector* has the same number of transmissions

B. Bursty arrival process

In Section III-A, we assumed small Poisson arrivals which allowed us to consider a truncated state space. Clearly, bursty arrival processes will not satisfy this condition. For large sized bursts, there is much to be gained by a splitting approach [16]. In this section, we reformulate the reservation problem for bursty arrivals, so that the contending nodes successively split into different classes based on their prior transmission history.

By assumption, the AP knows the size of the arrival burst, N . Thus, in the interval of time between two successive reservation frames, there will be N new arrivals which will make reservation attempts. *Theorem 1* allows us to consider this burst independently of other contending nodes in the system.

To illustrate our formulation, consider as an example, a burst arrival of $N = 6$ nodes in the interval between two reservation frames with the reservation deadline being $A = 3$ attempts. The question we want to answer is: how many minislots K_1 should the AP allocate to this burst class in the next reservation frame, to minimize the expected total number of minislots that will be allocated to this burst class over $A = 3$ frames to achieve a eventual probability of failed reservation, e.g. $P_{fail} < 0.01$.

Let $g(N, A, P_{fail})$ be the function that specifies the minimum expected number of total minislots required for a burst of size N to make reservations with a failure probability $\leq P_{fail}$ in A attempts, and $f(K_1)$ denote the expected total number of minislots allocated to the burst class over A attempts, given that K_1 minislots are allocated in next frame and an “optimum” allocation is used for subsequent frames.

$$g(N, A, P_{fail}) \equiv \min_{K_1} f(K_1) \quad (4)$$

A brute force procedure to find the optimal value of K_1 would be to search through a set of feasible values of K_1 , $K_1 = 1, 2, 3, \dots$ as suggested by Equation 4. For any fixed value of K_1 , the N nodes each use a optimum *transmission policy* computed as outlined in III-A. The result of this transmission attempt would be a *partitioning*: $\bar{N} \equiv [N_1, N_2, \dots, N_{K_1}]$, where N_{succ} is the size of the subset of the N contending nodes that made a successful reservation and N_i is the size of the subset of nodes involved in a collision in minislot i , $i \leq K_1$. In our example with $N = 6$, one possible partitioning is: $\bar{N} = [0, 2, 3]$ with $N_{succ} = 1$ and $K_1 = 3$ minislots.³

Any two partitionings that are just permutations of each other are mathematically identical. Therefore, we can identify a set of distinct partitionings and a probability distribution over

the set.⁴ Thus we have:

$$f(K_1) = K_1 + \sum_j P(\bar{N}^j) \sum_k g(N_k^j, A - 1, P_{fail}) \quad (5)$$

where $\bar{N}^j = [N_1^j, N_2^j, \dots, N_{K_1}^j]$ and $P(\bar{N}^j)$ is the probability of generating partitioning \bar{N}^j using the optimum transmission policy corresponding to N and K_1 .

Equations 4 and 5 define the dynamic programming procedure used to solve the bursty arrival problem. The first stage of the procedure computes $\{\bar{N}^j\}$ and their corresponding probabilities for a given value of K_1 . In the second stage, we repeat the same computation for each of the different collision classes for each partitioning with non-negligible probability. Finally, the procedure terminates with the last deadline state, $A = 1$, in which case the optimum transmission policy can be computed as in Section III-A. Note that for a given P_{fail} , it is possible to precompute the optimum K_i values for the feasible set of values of N and A , and this represents a complete solution to the optimization problem.

For a $P_{fail} = 0.04$, Figure 4 shows the total required expected reservation bandwidth (in terms of the number of total minislots) plotted against the size of the initial burst for different values of A . We make a few observations from these results.

- 1) *The required reservation bandwidth varies linearly with the burst size.* In other words, roughly twice the number of minislots is required to handle a burst of size 10 compared to a burst of size 5, given that the probability of failure is the same in both cases.
- 2) *Allowing more than about 4 total attempts does not significantly impact the total required reservation bandwidth.* This is demonstrated by Figure 4 where the curves seem to bunch closer for increasing values of A .
- 3) *It is optimal to allocate a small number of minislots until the last reservation attempt.*

Remark: The reservation bandwidth plotted in Figure 4 is an average quantity. The actual number of minislots required is a random variable which depends on the actual collisions observed. A typical spread is shown by the histogram in Figure 5, which is obtained by repeated Monte-Carlo simulation runs for a burst size of $N = 11$ nodes with $A = 4$ and required $P_{fail} < 0.04$. From the dynamic programming analysis, the expected value of number of required minislots for this P_{fail} requirement was approximately 34.5 minislots. The spread shown in Figure 5 has average of about 36.5. Data from the simulations also show that 89% of all collisions involved 2 nodes, 10% involved 3 nodes and $< 1\%$ involved more than 3 nodes. If the AP assumes each collision to be a 2 node collision and computes the subsequent transmission probabilities on this basis, the system would require a larger number of minislots to achieve the same P_{fail} . In our example, the resulting expected

³Depending on the optimum *transmission policy*, it is possible that the same node is involved in a collision in more than one minislot; in this case, we just assume that the node arbitrarily picks a “collision class” to associate with.

⁴In our implementation of the dynamic programming procedure, we used Monte-Carlo simulations to compute the different partitionings, and their probabilities are directly estimated from their relative frequencies. This method has the advantage of being intuitive, and of automatically filtering out very low probability partitionings.

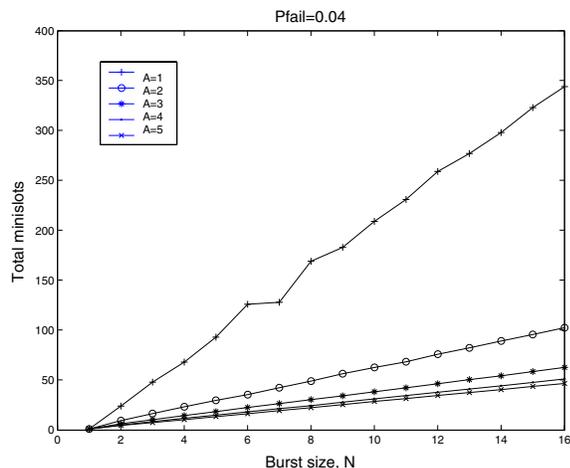


Fig. 4. Total reservation minislots against burst size

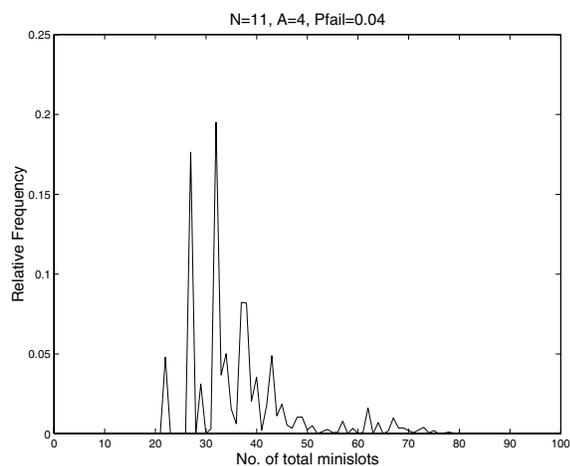


Fig. 5. Probability distribution of total minislots requirement

total minislots requirement is 42.5, an increase of about 15%. Although, in Section II, we assumed that the AP can estimate collision multiplicities, we find that the reservation scheme works reasonably well even without this information.

In Section II, we estimated the expected arrival rates for a pseudocellular network with mobiles moving at vehicular speeds to be about $\lambda \approx 0.2$ per reservation frame. Such an average arrival rate could be due to 1 arrival every 5 frames or a burst of 10 arrivals every 50 frames. Clearly, the second event is the worst case. For a probability of failure, $P_{fail} = 0.04$, we found that with $A = 4$, the total number of minislots required was $K_{total} \approx 40$. Assuming each reservation request required a 100 byte packet and neglecting Inter-Frame Spacings between two successive minislots, this represents a bandwidth requirement of 32,000 bits every 50 frames, i.e. an average bandwidth requirement of 13 kbps. For smaller values of P_{fail} , this will certainly increase, but it is clear that the overhead due to reservation requests is within manageable limits, so the AP will be able to allocate the required number of minislots without impacting the ongoing connections. This number is

consistent with our assumptions in Section II, and shows the basic feasibility of the WLAN pseudocellular network.

IV. CONCLUSION

We have shown that real-time traffic at vehicular speeds can be sustained in a pseudocellular network with small WLAN-type pseudocells, using polling or time division based resource sharing mechanisms, along with ALOHA-like reservation schemes optimized to provide probabilistic deadline guarantees. The methods we propose can be implemented using software upgrades alone, although they may be more efficient to implement in firmware. In addition to implementation of these methods, which focus on communication between a mobile and the nearest AP, an important problem for future research is to devise methods for efficient inter-AP connectivity for supporting real-time traffic in plug-and-play pseudocellular deployments.

REFERENCES

- [1] J. Sobrinho and A. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," *Bell Labs Technical Journal*, Autumn 1996.
- [2] Q. Ni, L. Romdhani, T. Turletti, and I. Aad, "QoS Issues and Enhancements for IEEE 802.11 Wireless LAN," <ftp://ftp-sop.inria.fr/pub/rapports/RR-4612.ps.gz>.
- [3] B. Crow, I. Widjaja, J. G. Kim, and P. Sakai, "IEEE 802.11 wireless local area networks," *IEEE Communications Magazine*, vol. 35, pp. 116–126, 1997.
- [4] IEEE 802.11 Working Group, "Wireless LAN Medium Access Control (MAC) And Physical Layer (PHY) Specifications: Higher-speed Physical Layer Extension In The 2.4 GHz Band," *IEEE standard document*, 1999.
- [5] "Packet Scheduling and QoS for Wireless Networks," <http://frottle.sourceforge.net/>.
- [6] D. Bertsekas and R. Gallager, *Data Networks*. Prentice Hall, 1987.
- [7] IEEE 802.14 Cable TV Protocol Working Group, "Formal MAC Proposals," November 1995.
- [8] D. Sala and J.O. Limb, "Comparison of contention resolution algorithms for a cable modem MAC protocol," in *Proceedings of the International Zurich Seminar on Broadband Communications*, pp. 83–90, 1998.
- [9] N. Abramson, "Development of the ALOHANET," *IEEE Transactions on Information Theory*, vol. 31, pp. 119–123, May 1985.
- [10] Y. Birk and Y. Keren, "Judicious use of redundant transmissions in multichannel ALOHA networks with deadlines," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 257–269, February 1999.
- [11] E. Wong and T. Yum, "The optimal multicopy ALOHA," *IEEE Journal on Automatic Control*, vol. 39, pp. 1233–1236, June 1994.
- [12] D. Vazquez-Cortizo, C. Blondia, and J. Garcia, "FS-ALOHA++, a collision resolution algorithm with QoS support for the contention channel in multiservices wireless LAN," in *Global Telecommunications Conference, 1999 (GLOBECOM 99)*, vol. 5, pp. 2773–2777, 1999.
- [13] K. Bruvold and U. Madhow, "Adaptive multi-user detection for mobile-centric fast handoffs in pseudocellular wireless networks," in *Proc. 58th IEEE Vehicular Technology Conference (VTC Fall'03)*, October 2003.
- [14] W. Goralsky and M. Kolon, *IP Telephony*. McGraw-Hill Inc., 1999.
- [15] D. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Athena Scientific, 1995.
- [16] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall Inc., 1992.