On the use of Multiple Working Points in Multichannel ALOHA with Deadlines*

Dror Baron and Yitzhak Birk
Electrical Engineering Department
Technion — Israel Inst. of Technology
Haifa 32000, Israel
birk@ee, drorb@tx.technion.ac.il

Abstract

This paper addresses the problem of maximizing the capacity of multichannel Slotted ALOHA networks subject to a deadline and a permissible probability of exceeding it. A previous paper proposed to transmit a non-decreasing number of copies of a message in successive rounds until success or deadline. This yielded a low probability of failure due to the large maximum number of copies per message, with only minimal "pollution" due to the small mean number of copies. In this paper, we examine another way of implementing variable resource expenditure in different rounds: the channels are partitioned into groups, one for each round (until the deadline), and the channels used by later rounds are operated with lower offered loads. These Multiple Working Point (MWP) policies are shown to attain a lower capacity than the optimal Multicopy (MC) scheme. Combining the two to form an MC-MWP scheme slightly improves capacity over MC-SWP. The SC-MWP approach can be more attractive when using a single transmitter per station because, unlike MC, transmission time is not prolonged. Therefore, as the trend from high orbit satellites to networks with lower propagation delays continues, Multiple Working Point policies should become of more interest.

Keywords: Multichannel ALOHA, satellite networks, deadline, multiple working points.

1 Introduction

ALOHA [1] is the simplest access scheme because it does not require channel sensing or collision detection, but performs worse than more elaborate schemes when those are practical. An important use of ALOHA at present is by satellite ground stations, because the long propagation delay precludes timely channel sensing. ALOHA is used as the primary access scheme for short messages, and in order to reserve channels for long ones [2].

Fig. 1 depicts a typical satellite-based ALOHA network. The *stations* transmit data in globally synchronized time slots over contention uplink channels (dashed lines). Successful reception by the *hub* is acknowledged by it immediately over contention-free downlinks

*This work was supported in part by the Information Superhighway In Space consortium, administered by the office of the Chief Scientist of the Israeli Ministry of Industry and Trade.

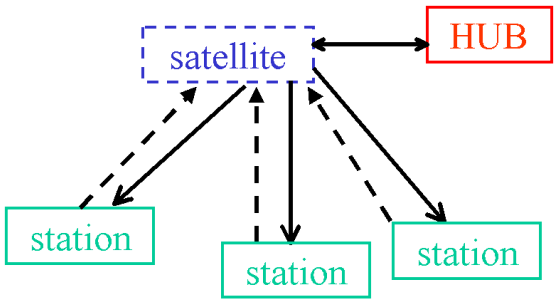

Figure 1: A typical hub-based satellite network.

(solid lines). The hub can be terrestrial or part of the satellite. If several simultaneous transmissions occur, none of them succeed. Stations can only learn about a collision through the absence of an acknowledgment. The time from the beginning of a transmission until the time by which an ACK for it must be received (or else it is considered to have collided) is referred to as a *round*. Unlike slots, which must be synchronized among the stations, a round is "private" and requires no coordination. A station retransmits packets until they succeed or until a deadline is exceeded. The typical duration of a round is up to several tens of slots.

In a single-channel ALOHA network, the retransmission delay (upon not receiving an ACK) must be randomized to prevent definite repeated collisions. To improve stability, a station must moreover increase the mean back-off time in later rounds. Current satellite networks employ as many as hundreds of channels. When operated with ALOHA, e.g., for small transactions, a station picks a channel at random for each transmission. The hub can receive concurrently over all channels. The randomized retransmission delay is replaced with immediate retransmission over a randomly chosen channel.

Over the years, the bulk of the research on ALOHA and related reservation schemes, e.g. [3], concerned maximizing capacity. Some attention was given to delay-throughput trade-offs and other performance measures. The advent of multichannel ALOHA networks has given rise to the use of redundant transmissions for performance improvement. For example, [4] studies *Multicopy ALOHA* (MC), whereby a station transmits several *copies* of a packet in each round, as a way of improving delay–throughput performance. (We refer to this as "redundancy" because, unlike retransmission upon failure, some of the transmissions may not be required.)

Recently, Birk and Keren [5] proposed an optimization problem that reflects both intuitive user requirements and the desires of network designers: maximization of capacity subject to a deadline and a permissible probability of exceeding it. They proposed a *non-stationary Multicopy* (MC) transmission policy, whereby a station transmits a monotonically non-decreasing number of copies in successive rounds until successful reception or deadline. Dynamic programming was used to optimize the transmission sequence, resulting in a substantial increase in capacity relative to that attainable with classical ALOHA or even with (fixed) MC ALOHA [4]. The advantage is more pronounced for stricter constraints. They moreover adapted the optimized scheme to the practical situation wherein a station only has a single transmitter. This was done by transmitting a burst of copies in successive slots over randomly chosen channels, and then waiting for ACKs for all of them before proceeding to the next round. This technique, dubbed *Round*

Stretching, was shown to achieve similar capacities to the multi-transmitter scheme in most situations. Fig. 2 illustrates the idea. Note that, for any given deadline, Round Stretching may reduce the permissible number of rounds.

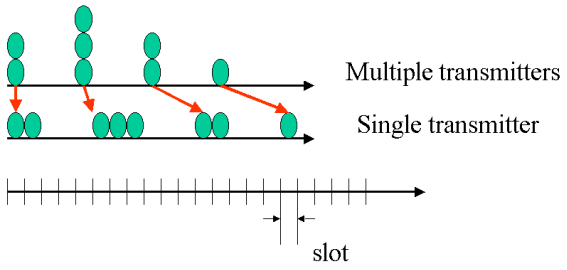

Figure 2: Round Stretching.

One can use *pure* MC policies, whereby the number of copies transmitted in any given round is deterministic (albeit not the same for all rounds), or *impure* policies whereby it is randomized. This idea is studied in [6] in the context of optimizing the throughput–delay trade-off with MC ALOHA. Preliminary research results [8] suggest that an impure variant of the replication-based scheme of [5] produces an insignificant increase in capacity.

Multislot messages were first considered in [5] for single-round transmissions. A multi-round approach is developed in [8]. For a K-slot message, redundant single-slot fragments are computed using block erasure-correcting codes, such that any K fragments suffice for message reception. With the *Multiround Coding* scheme, an optimized number of fragments are transmitted in each round until K are received or the deadline is reached. Even with very strict constraints, capacities that approach the $1/e$ limit are attained. The *Coding–Reservation* scheme, also proposed in [8], raises capacity above $1/e$ by using the foregoing fragment transmissions to also request contention-free channels, which are granted once some fragment(s) are received prior to the deadline and used for the remaining required fragments.

The key idea in the replication-based scheme of [5], which is employed in this paper as well, is to permit a large maximum channel-resource expenditure per message while keeping the mean expenditure low. This is done by being more "wasteful" in the later rounds, which are less likely to even take place. By so doing, the probability of failure can be made very low (because a message fails only after the maximum has been spent on it) without giving up much capacity. In [5], the expenditure manifested itself as speculative transmission of multiple copies in late rounds.

In this paper we propose and study an alternative way of controlling the resource expenditure: the channels are partitioned into groups, one per round, with lower offered loads in the channels used for later rounds. (The number of channel groups equals the number of permissible transmission rounds until deadline.) This approach is dubbed *Multiple Working Points* (MWP). We begin by comparing the SC-MWP scheme (single copy per round) with the MC-SWP scheme of [5]. Then, the methods are combined into MC-MWP. Only single-slot messages are considered.

The remainder of the paper is organized as follows. In Section 2, we present the network model that is subsequently used for performance analysis, derive some prelimi-

nary mathematical relations for use in later sections, and define the design space of our problem. Section 3 proves that each round should best a single working point and a deterministic number of copies. Section 4 provides a general mathematical analysis of MC MWP policies, Section 5 presents performance results, and Section 6 offers concluding remarks.

2 Network model and preliminaries

2.1 Model and definitions

The network comprises ground stations that transmit single-slot messages over randomly chosen channels. A hub monitors all channels and ACKs all successful receptions. The lack of an ACK when it is expected indicates a collision. A station transmits in rounds, waiting for the results of one round before continuing to the next, until the deadline; then, an as-yet unreceived message is declared lost. (We will consider very small permissible loss probabilities. Therefore, "lost" messages may be reissued with negligible effect on performance.)

We assume an infinite number of stations and a large number of channels. The number of transmissions over any given contention channel in any given time slot is a Poisson random variable, independent from slot to slot and from channel to channel. Thus, the probability of collision of packet is only a function of the offered load on the channel. This is an approximation, but the large number of channels along with the randomization make it close, and causes the degradation when used in finite networks to be graceful. Finally, the inaccuracies are unlikely to distort the comparison among schemes.

ALOHA can be bistable in certain regions. However, especially in hub-based multichannel networks, it is possible to employ algorithms that detect such situations and "push" the network into the "good" stable point. The analysis in this paper applies to stable operation.

A user-specified deadline is expressed in time units. For facility of exposition, we define this to be the time from the first transmission until the time of the latest transmission that would still arrive by the deadline. With fixed-size slots, we use D_s to express the deadline in slots. For rounds of fixed duration, we use D_r to denote the maximum permissible number of rounds. P_e denotes the permissible probability of missing the deadline.

When Round Stretching [5] is used, let T_A denote the number of slots from single-slot transmission until ACK (assuming success) or until the next transmission may take place (assuming collision). Then,

$$D_s = (D_r - 1)T_A + N_{max}, \tag{1}$$

where N_{max} is the maximum total number of transmissions of a message until its deadline. When $T_A \gg 1$, D_r is not affected much by N_{max}, and Round Stretching hardly changes performance. For small T_A, the effect varies.

Channel Capacity. Because messages may be dropped, albeit with a low probability, a distinction was made in [5] between the generation rate of messages, S_g, and the throughput S. Specifically, $S = (1 - P_e)S_g$. Derivation of channel capacity is complicated in MWP networks by the fact that several groups of channels are used. If the mean traffic on a set of contention channels with an offered load G is n transmissions per slot, then the required number of channel slots is $\frac{n}{G}$. ("Channel slots" is a measure of channel

resources, not delay.) We therefore derive both the capacity S and the generation rate S_g by examining the mean number of channel slots consumed by a message (successful or generated for S and S_g respectively). Consequently, we can again write $S = (1 - P_e)S_g$.

2.2 Useful relations

For pure MC-SWP policies [5],

$$S_g = \frac{G}{E[N]}, \tag{2}$$

where $E[N]$ denotes the expected number of transmitted copies per message until success or deadline. Channel capacity is thus

$$S = S_g(1 - P_e) = \frac{G(1 - P_e)}{E[N]}. \tag{3}$$

The total number of copies transmitted per message is $N = \sum_i n_i \leq \sum_{i=1}^{D_r} n_i \leq N_{max}$, where n_i denotes the number of copies transmitted in round i. The probability of collision is $P_c = 1 - e^{-G}$. Since $P[\text{reach round } i] = (P_c)^{\sum_{j=1}^{i-1} n_j}$, the expected total number of copies per message is

$$E[N] = n_1 + \sum_{i=2}^{D_r} n_i (P_c)^{\sum_{j=1}^{i-1} n_j}. \tag{4}$$

2.3 Design Space

The design space for single-slot messages has several dimensions: single/multiple copies per round; single/multiple working points; stationary/non-stationary; and pure/impure. Stationary policies do the same thing in every round, whereas non-stationary ones are round-dependent. Pure policies are deterministic, whereas impure ones are probabilistic, e.g. choose among several working points or among several numbers of copies in a given round. (Selection of a channel among those in the same group does not constitute impurity.)

The remainder of the paper is organized as follows. In Section 3, we prove that with MWP policies, each round should best use a pure SWP policy. This, combined with simulation results whereby ***, allows us to focus on pure policies. Section 4 provides performance analysis for MC-MWP case, of which SC-MWP is a special case. Section sec:mwpresults*** and Section sec:mwpconclusions offers concluding remarks.

3 Optimality of a single working point per round

Impurity of an MC-MWP policy can entail a probabilistic choice of the nubmer of copies in any given round (with a possibly different mean for each round), as well as a probabilistic choice among several working points. Numerical results have shown that, with multi-round MC-SWP policies, a probabilistic number of copies can increase capacity, but the increase is minute; also, pure ones appear to be optimal for a single round [**?**]. In view of this, we assume a deterministic number of copies in each round, denoted n_i.

In the remainder of this section, we prove that it is best to use a single working point for each round.

Recalling the assumptions that were made, and that the fate of a transmitted copy is only influenced by the offered load (working point) of the channel that it is using, let us consider a single round. We initially assume that it is allocated two sets of channels, operated at different working points, and show that it would have been better to use a single (different) working point.

Theorem 1 *SWP policies are optimal among single-round policies that transmit a deterministic number of copies at each of the WPs they use.*

Proof: We will prove that replacing two copies transmitted at two different WPs with two copies transmitted at a single WP, such that the probability that both fail is unchanged, reduces required channel resources. Since only SWP policies don't have mergable copies, this implies that only SWP policies can be optimal.

Consider two copies transmitted at two WPs with channel error probabilities P_{c_1} and P_{c_2}. The probability both copies colliding is $P_{c_1} \cdot P_{c_2}$.

Suppose WP_m $(m = 1, 2)$ operates at an offered load of G_m copies per channel per slot. If the combined mean traffic over all channels operating at WP_m is n_m copies per slot, the requred number of channels is

$$W_m = \frac{n_m}{G_m}. \tag{5}$$

Therefore, the total number of channels required for the two working points in "support" of the transmission of a single copy at each of them is

$$W = \frac{1}{G_1} + \frac{1}{G_2}. \tag{6}$$

The same probability of error can be obtained by transmitting two copies at a single WP with collision probability

$$\tilde{P}_c = \sqrt{P_{c_1} P_{c_2}}. \tag{7}$$

According to (5),

$$\tilde{W} = \frac{2}{\tilde{G}} \tag{8}$$

channels are required. It suffices to show that $\tilde{W} \leq W$. For the detailed proof, see [7]. ⊔

4 Capacity of pure MC-MWP schemes

$$P_e = \prod_{i=1}^{D_r} (P_{c_i})^{n_i}. \tag{9}$$

The generation rate at WP i is

$$S_{g_i} = \frac{G_i}{E[N_i]} = \frac{G_i}{n_i}. \tag{10}$$

The number of channels necessary for later rounds is affected by the amount of messages entering those rounds, which is lower if "cleaner" WPs are used in earlier rounds. In order to derive the network capacity (mean successful messages per channel slot), we must calculate the number of channels necessary for each WP, followed by the throughput obtained in each channel. Summing the throughputs and dividing by the total number of channels yields the normalized (per channel) capacity.

Consider W_1 channels used for WP 1 in the first round. The generation rate of messages in the network is $S_{g_1}W_1$. The rate of messages entering round 2 is $S_{g_1}W_1(P_{c_1})^{n_1}$. However, it is also equal to $S_{g_2}W_2$, so

$$W_2 = W_1 \frac{S_{g_1}}{S_{g_2}}(P_{c_1})^{n_1}. \tag{11}$$

Similarly, $S_{g_1}W_1 \prod_{k=1}^{i-1}(P_{c_k})^{n_k}$ messages enter round i, and

$$W_i = W_1 \frac{S_{g_1}}{S_{g_i}} \prod_{k=1}^{i-1}(P_{c_k})^{n_k}, i \geq 2. \tag{12}$$

The network generation rate is

$$S_g = \frac{W_1 S_{g_1}}{W_1 + \sum_{i=2}^{D_r} W_1 \frac{S_{g_1}}{S_{g_i}} \prod_{k=1}^{i-1}(P_{c_k})^{n_k}}, \tag{13}$$

and, according to (**??**),(10), and purity in round i,

$$\begin{aligned} \frac{1}{S_g} &= \frac{1}{S_{g_1}} + \sum_{i=2}^{D_r} \frac{1}{S_{g_i}} \prod_{k=1}^{i-1}(P_{c_k})^{n_k} \\ &= \frac{n_1}{G_1} + \sum_{i=2}^{D_r} \frac{n_i}{G_i} \prod_{k=1}^{i-1}(1-e^{-G_k})^{n_k}, \end{aligned} \tag{14}$$

and the capacity is $S = S_g(1 - P_e)$.

5 Numerical results

In order to compare the performance of MWP policies with SWP policies, a computer program that, given (n_i) and P_e, optimizes $\{G_i\}$ according to (14), was written. If necessary, an external loop on (n_i) is performed, and the best result is picked.

This section is organized as follows. First, performance of SC MWP policies will be examined, and a comparison between SC MWP and MC SWP mechanisms will explain performance differences. Then, we elaborate to MC MWP policies, and their capacity will be shown to be slightly better than optimal MC SWP policies [5]. Finally, Round Stretching in MWP policies will be studied.

5.1 SC policies

Table 1 shows the performance of SC MWP and SC SWP policies for several delay constraints. The use of multiple WPs provides a major performance boost.

Table 1: The capacity of MWP and SWP policies.

D_r	P_e	SWP			MWP		
		SC	Optimal MC		SC	Optimal MC	
		S	(n_i)	S	S	(n_i)	S
3	10^{-2}	0.190	1,2,4	0.279	0.233	1,2,4	0.281
	10^{-3}	0.095	2,3,7	0.247	0.158	1,2,6	0.248
	10^{-4}	0.045	2,3,10	0.233	0.110	2,3,9	0.234
5	10^{-2}	0.306	1,1,1,2,3	0.340	0.335	1,1,1,2,2	0.342
	10^{-3}	0.217	1,1,1,2,5	0.321	0.296	1,1,1,2,5	0.324
	10^{-4}	0.145	1,1,2,3,8	0.313	0.264	1,1,2,3,7	0.314

5.2 Comparing mechanisms

The added dimensions of freedom of multiple WPs are certainly Beneficial. Therefore, SC MWP is better than SC SWP. However, optimal MC SWP policies [5] are even better. Examining the differences between the MC SWP and SC MWP mechanisms can provide insight.

In optimal MC SWP policies [5], the probability that the message is successfully transmitted in later rounds is increased by increasing the number of copies used in those rounds. The probability of failing to meet the deadline, P_e, decays exponentially in N_{max}, the maximal total number of copies transmitted.

$$P_e = (P_c)^{N_{max}}. \tag{15}$$

Therefore, given some WP, N_{max} is logarithmic in P_e, so

$$N_{max} = \frac{\ln P_e}{\ln P_c}. \tag{16}$$

However, the message does not always utilize all the rounds, so the expected total number of copies transmitted per message, $E[N]$, grows less than logarithmically in P_e. The cost, in terms of channels, needed to maintain a low error probability, is not very high.

In SC MWP policies, the probability that the message is successfully received is increased for late rounds by maintaining "clean" WPs for those rounds. The offered load on those channels is (**??**)

$$G \approx 1 - e^{-G} = P_c, \quad G \ll 1. \tag{17}$$

Therefore, according to (5) the number of channels required for each "late" round is roughly inversely proportional to the probability for channel collision in that round. The cost, in terms of channels, needed to maintain a low error probability, is quite significant.

5.3 Optimal policies

When deadlines are added, MC SWP policies provide major capacity improvements over classical ALOHA [5]. When multiple WPs are added, there is a small performance improvement for the same sequence (n_i). Using another sequence (n_i) is sometimes even better. Therefore, an external loop on (n_i) is needed in order to find the optimal MC MWP policy.

Table 1 shows the performance of optimal MC MWP and optimal MC SWP policies, for several delay constraints. The improvements in capacity are below 1%. The conclusion is that if capacity is the main design goal, using MWP policies might not be worth the added implementation complexity.

5.4 MWP Round Stretching

In Table 1, even when SC policies were used, the performance of MWP policies was reasonable. When Round Stretching is considered for (possibly MC) MWP policies, a major concern is the trade-off between adding an additional round, or increasing N_{max}. Numerical results suggest that for MWP policies, it is usually better to use as many rounds as possible, although this keeps N_{max} small.

Table 2: MWP slot savings vs. optimal MC SWP policies [5].

D_r	P_e	SWP			MWP			Savings
		(n_i)	N_{max}	S	(n_i)	N_{max}	S	
3	10^{-2}	1,2,4	7	0.279	1,2,3	6	0.280	1
	10^{-3}	2,3,7	12	0.247	1,2,5	8	0.247	4
	10^{-4}	2,3,10	15	0.233	2,3,7	12	0.233	3
5	10^{-2}	1,1,1,2,3	8	0.340	1,1,1,1,2	6	0.341	2
	10^{-3}	1,1,1,2,5	10	0.321	1,1,1,2,3	8	0.323	2
	10^{-4}	1,1,2,3,8	15	0.313	1,1,2,2,5	11	0.313	4

Since MC MWP policies have slightly higher capacity than MC SWP policies, we decided to check how many slots of delay can be saved using MWP policies, while attaining at least the same capacity as the optimal MC SWP policy [5]. We held D_r constant, and checked how much N_{max} can be decreased. Table 2 shows that MWP policies can provide a significant savings in N_{max}. When stricter delay constraints are used, N_{max} rises, as does the savings in slots.

Fig. 3 depicts channel capacity with Round Stretching for MWP policies using the (P_e, D_s) constraint. Results for Classical ALOHA and SWP policies [5] are shown for reference. The figure also depicts a MWP policy with an unlimited number of transmitters per station. The conclusions are as follows:

- For large D_s, channel capacity approaches $1/e$, the upper bound on capacity with Slotted ALOHA.

- For any given scheme, capacity increases with an increase in D_r. With Round Stretching, however, especially for values of D_s that barely permit another round, one must decide whether to increase D_r at the cost of significantly reducing N_{max} or stay with one fewer round and slightly increase N_{max}. The result of optimization is that, as D_s is increased and permits an additional round, the channel capacity with multiple transmitters rises immediately, whereas that with Round Stretching stays flat until such value of D_s for which an increase in D_r is warranted. Then, capacity rises sharply and eventually comes close to that with multiple transmitters per station. The MWP policies, due to their ability to use "clean" last rounds, cope well with constraints on N_{max}. Therefore, when optimized, they elect to use an

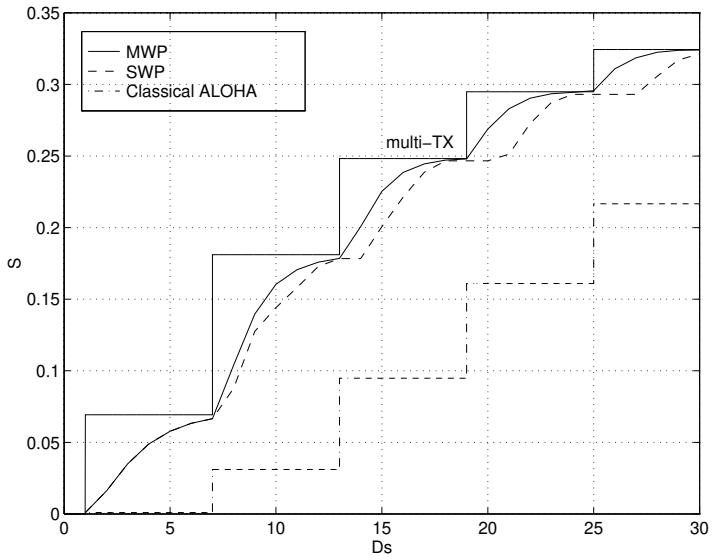

Figure 3: Channel capacity of MWP and SWP policies with Round Stretching. $P_e = 10^{-3}$; $T_A = 5$.

additional round earlier (already at smaller values of D_s) than SWP policies do. For this reason, MWP policies can save several slots of delay in Round Stretching.

- In Fig. 3, MWP and SWP policies have identical performance up to $D_s = 7$, because a single round is used and the same (single) WP is thus chosen.

6 Conclusions

This paper focused on the maximization of capacity for single-slot messages in multi-channel Slotted ALOHA networks. It explored the use of different working points in different rounds as a means of implementing non-stationary expenditure of network resources in order to achieve low probabilities of failure while holding down the mean per-message resource expenditure. Through numerical results as well as some analytical insight, this this Multiple-Working-Point approach was shown to be generally inferior to controlling the number of copies per round. (Nonetheless, it is significantly advantagous over the conventional SC-SWP approach.) An MC-MWP hybrid offers only a slight advantage when a station is equipped with multiple transmitters, but this advantage increases in the case of a single transmitter and round stretching, especially when the permissible delay is small and the permissible probability of failure is small. As the trend from high orbit satellites to networks with lower propagation delays (and thus fewer slots per round) continues, MWP policies, should become of greater interest.

One direction for future research is the use of MWP policies for multislot messages. The combination of *Coding–Reservation* schemes [8] that provide impressive capacity gains, with MWP policies that do well for Round Stretching, is certainly of interest. Another direction for future research involves multiple service categories. Several classes of messages requiring different qualities of service (different constraints) can be allocated several sets of policies, each using some set of WPs. However, joint optimization of the problem might be better.

Finally, we note that the results of this paper serve as yet another example of the benefits gained from the judicious use of redundancy for performance enhancement.

References

[1] N. Abramson,"Development of the ALOHANET", *IEEE IT*, vol. 31, no. 3, pp. 119-123, Mar. 1985.

[2] R. Rom and M. Sidi, *Multiple Access Protocols*. New York; Springer-Verlag, 1990.

[3] S. S. Lam,"Packet Broadcast Networks - A Performance Analysis of the R-ALOHA Protocol", *IEEE Trans. Computers*, vol. C-29, no. 7, pp. 596-603, July 1980.

[4] E. W. M. Wong and T. S. P. Yum,"The Optimal Multicopy Aloha", *IEEE Trans. Auto. Control*, vol. 39, no. 6, pp. 1233-1236, June 1994.

[5] Y. Birk and Y. Keren,"Judicious Use of Redundant Transmissions in Multi-Channel ALOHA Networks with Deadlines", *IEEE JSAC*, vol. 17, no. 2, pp. 257-269, Feb. 1999.

[6] Y. W. Leung,"Generalised multicopy ALOHA", *Electronics Letters*, vol. 31, no. 2, pp. 82-83, Jan. 1995.

[7] D. Baron and Y. Birk, "Multiple working points in multichannel slotted ALOHA networks with deadlines", EE Tech report No. 1220 (also CC-pub 292, Technion, Haifa Israel, Sep. 1999.

[8] D. Baron and Y. Birk, "Coding and Coding–Reservation Schemes for Multislot Messages in Multichannel ALOHA with Deadlines", EE Tech report No. 1224, (also CC-pub 293), Technion, Haifa Israel, Sep. 1999.