Capacity Maximization in Multichannel Slotted ALOHA with Deadlines – an Overview*

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

Rev. June 11, 2000

Abstract

Slotted multichannel ALOHA is the access scheme of choice for sending short messages and for reserving channels to be used by longer ones in many satellite-based networks. For these applications, maximization of capacity subject to meeting a specified deadline with a specified (high) probability jointly captures true, intuitive user requirements and the desire to maximize system cost-effectiveness. This report addresses deadline-constrained capacity maximization. A key idea is to achieve a low probability of missing the deadline by permitting a large maximum resource expenditure per message, while holding the mean expenditure low in order to minimize “pollution”. This is achieved through the judicious exploitation of redundancy, e.g., transmission of an increasing number of copies per round as the deadline approaches, or by the use of different working points for channels employed in different rounds. The judicious exploitation of redundancy substantially increases the delay-constrained capacity and is moreover practical and even power-efficient. This report provides a road map of our recent work on this topic, and refers the interested reader to other publications for the details.

Keywords: Multichannel ALOHA, satellite networks, deadline, multiple working points, redundancy, multi-copy.

*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. Electrical Engr. TR 1248 (also CCIT TR 314), June 2000, Technion.
†Presently with the ECE Dept., Univ. of Illinois at Urbana-Champaign, dbaron@ipf.uiuc.edu.
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. It is used as the primary access scheme for short messages, and in order to reserve channels for long ones [2]. Another application in which ALOHA is widely used is cellular networks, wherein the control uplink channels from the cellular phones to base stations are multiple access. A future application for ALOHA may be transmission of short messages over high speed point-to-multipoint terrestrial wireless networks: a central base station could collect transmissions from a large number of users using shared bandwidth. Since propagation delay for terrestrial stations 10 kilometers apart is on the order of 0.1 milliseconds, even present data rates permit transmission of many thousands of bits during this time, precluding timely channel sensing.

Figure 1: A typical hub-based satellite network.

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 (solid lines). The hub can be terrestrial or in space. If several simultaneous transmissions occur over a given channel, they all fail. 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, retransmission delay must be randomized in order to prevent definite repeated collisions [3]. 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 the transmission of multiple copies per round as “redundancy” because, unlike retransmission upon failure, some of the transmissions may not be required.

Virtually all current applications of ALOHA entail the transmission of single packets or ones that occupy a very small number of time slots, be it for short transactions or to reserve channel resources for the transmission of large amounts of data. Also, the user is typically charged per actual traffic, while the system owner pays for bandwidth (channel) resources. From a user perspective, the key performance criterion is delay, and it is most naturally expressed as a constraint (e.g., a deadline along with a permissible probability of exceeding it). From the system owner’s perspective, capacity maximization is the main design goal. Our interest is thus in the maximization of network capacity subject to a requirement that a given deadline be met with a certain (very high) probability.

2 Maximization of delay-constrained capacity

An important implication of the chosen performance measure is that there is no reward for reducing mean delay. One must merely meet the deadline with the required probability. Thus, it is acceptable to increase the mean delay for purposes such as reduction in offered load and thus collision probability, so long as this does not reduce the probability of meeting the deadline.

Birk and Keren [5] 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.

Consider, for example, a situation wherein the deadline permits two rounds, the probability of collision of a packet is 0.1 (independent from packet to packet), and the permissible probability of failing to meet the deadline is 0.0001. One possible solution is to transmit four copies in the first round. The mean delay (given success) would be 1.0, and the mean channel resource spent per packet would be 4.0. Instead, consider transmitting a single copy in the first round and, if and only if it fails, transmitting three copies in the second round. The probability of failing to meet the deadline is unchanged, but the mean resource expenditure per packet drops to 1.3. One could thus increase the number of active clients (and thus capacity) three fold while maintaining the same probability of collision. The mean delay (given success) is only slightly higher, 1.1. Moreover, this does not matter!

The optimized non-stationary multicopy scheme was also adapted in [5] to the practi-
cal 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.

![Diagram of multiple transmitters and single transmitter](image)

**Figure 2:** Round Stretching.

The key idea underlying the replication-based scheme of [5] 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 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 [6], we propose and study an alternative way of controlling the resource expenditure: the channels are partitioned into groups, one per permissible transmission round, with lower offered loads on the channels used for later rounds. This approach is dubbed Multiple Working Points (MWP). We begin by comparing an MWP, single-copy-per-round scheme (SC-MWP) with the multi-copy, single-working-point (MC-SWP) scheme of [5]. Then, the methods are combined into MC-MWP. The conclusion is that the use of multiple working points is less effective than the use of multiple copies. Also, the optimal combination of the two increases capacity only slightly over multiple copies.

One can use pure multi-copy 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 [7] in the context of optimizing the throughput–delay trade-off with MC ALOHA. An impure variant of the replication-based scheme of [5] is studied in [8]. It achieves higher capacity than the optimal pure policy for multiple rounds, so optimal policies may be impure. Nonetheless, the advantage is very small. In the single round case, impurity appears to be of no benefit.

Multi-slot messages were first considered in [5] for single-round transmissions. A multi-round approach is developed in [9]. For a $K$-slot message, redundant single-slot fragments are constructed 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 [9], 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 are used for the remaining required fragments.

Yet another way of controlling the amount of channel resource allocated to a packet is transmission power. Birk and Revah have embarked on a study of the use of power capture for increasing delay-constrained capacity of multi-channel ALOHA. Preliminary results are positive. In fact, the advantage of power capture is much more pronounced for delay-constrained capacity than for unconstrained capacity.

### 3 Conclusions

Capacity maximization of multi-channel Slotted ALOHA networks for single-slot messages subject to a deadline and a permissible probability of failing to meet it is a goal that faithfully represents user requirements and designer goals for the current uses of ALOHA. A key idea, suggested in [5] and embodied in an optimized multi-copy scheme, is to achieve a small mean per-message resource expenditure while permitting a large maximum expenditure. This is instrumental in attaining a very low probability of missing the deadline without unnecessary “pollution” of the spectrum. We have explored the use of channels with different working points for different rounds both as an alternative method and as a complementary one. We have discovered that multiple working points are not as effective as multiple copies, and the combination of the two increases capacity only slightly. Also, we showed that an optimal non-deterministic (“impure”) multi-copy scheme can achieve a slightly higher capacity than the optimal deterministic (“pure” scheme). The use of multiple power levels with power capture is presently being studied.

For multi-slot messages, we extended the multi-copy scheme to the use of erasure correcting codes, and furthermore combined this with making reservations. This yields a significant capacity increase over the independent application of the optimal multi-copy scheme to each single-slot message fragment.

It is interesting to note that the capacity-maximization schemes that we have developed achieve their capacities with the minimum possible mean per-message channel-resource expenditure. Thus, they are also energy efficient, a desirable feature for mobile devices. An interesting problem for future research is energy minimization when operating below the delay-constrained capacity.

Finally, we note that the foregoing results serve as yet another example of the benefits gained from the judicious use of redundancy for performance enhancement. By deferring the expenditure of redundancy to the late rounds, we were able to attain a low probability of missing the deadline with only a small loss of capacity relative to unconstrained capacity.
References


