Announced Dynamic Access Probability protocol for next generation wireless networks

Z. NAOR [*]
H. LEVY [†]
School of Computer Science
Tel Aviv University, Tel Aviv 69978 Israel

October 21, 2002

Abstract

A multiple access protocol that is particularly suitable for cellular Internet access and satellite-based networks with on-board processing is developed in this paper. The basic idea is that when a user wishes to send a message, it transmits with probability p_{access} that depends on the load on the channel. Under conditions of low load, the probability p_{access} approaches 1, while at high load p_{access} is relatively low. This media access control protocol guarantees high channel utilization at high load, as well as low delay at low load periods. Using the statistical usage of the shared channel, the load is estimated with certain uncertainty. Our analysis shows that using the statistical usage of the shared channel, the optimal access probability can be well estimated for a broad class of load distribution patterns. In addition, we propose to use a central station to broadcast the value of p_{access} in networks with poor collision detection capability, or long feedback delay. The proposed method is particularly suitable for shared channels with poor collision detection capability, under conditions of bursty traffic and a large number of users. Examples for such channels are the reservation channel in satellite-based networks with on-board processing, and the control channel in cellular networks. Hence, the proposed method can be used for cellular Internet access and for accessing public satellite-based networks. *The broadcast mechanism that already exists in such networks can be used to inform the users the dynamic access probability.*

Keywords: wireless networks, multiple access, MAC

[*] `naorz@post.tau.ac.il`
[†] `hanoch@cs.tau.ac.il`

1 Introduction

Wireless networks are rapidly expanding. Future satellite-based networks will have to support a large number of users, and will have to provide services such as cellular Internet access and video-on-demand. The access channel in these networks is a shared channel with poor collision detection capability, and conditions of bursty traffic and a huge number of users. For this reason, the traditional channel access control protocols based on immediate collision detection capability, are not applicable on such networks. The goal of this study is to develop a multiple access protocol that is particularly suitable for such networks.

Networks in which many independent users share a common channel are widely spread. Examples for such networks are wireless networks and local area networks. A key issue in these networks is the allocation of a shared channel among many competing stations wishing to use it. There are three basic strategies for shared channel acquisition which are known in the literature: The first such strategy is that of contention protocols, as the ALOHA [1] and CSMA ([2], [16]) protocols. The second strategy includes collision-free methods, such as Time Division Multiplexing (TDM) ([2], [16]) and Frequency Division Multiplexing (FDM) ([2], [16]), while the third strategy is known as limited contention.

The basic idea of contention protocols is that users transmit whenever they have data to send. As a result, collisions are unavoidable. An example of a contention protocol is the ALOHA [1] and its many variants. Another example of contention is the carrier sense protocols: For carrier sense protocols, in order to reduce the likelihood of collision, users detect the state of the channel whenever they wish to transmit, but only send a message if no other user is transmitting. An example of a carrier sense protocol is the CSMA and its many variants ([2], [16]).

Collision-free protocols divide the users into disjoint groups, such that a collision is impossible. An example of a collision-free protocol is the Time Division Multiplexing (TDM) protocol ([2], [16]), in which each user has a dedicated time slot.

Another approach, which is a combination of the contention and the collision-free strategies, is the limited contention strategy. This strategy employs the contention protocol at periods of low loads, and a collision-free strategy at periods of high load. This is accomplished by dividing the stations into groups, which are not necessarily disjoint, such that at any given time, only one member of each group is permitted to transmit. The number of users in each group changes dynamically as a function of the load.

The contention protocols are preferable under conditions of light load, because of their low delay.

As the load increases, the likelihood of collisions becomes significant, and the channel efficiency of contention protocols at high load is very poor. For example, the channel utilization of the ALOHA protocol is very poor under high load. In order to overcome this problem, a variant of the ALOHA protocol was proposed, in which each user attempts to access the shared channel with some predefined *fixed* probability p, $0 < p < 1$. In comparison to the original or slotted ALOHA protocols, this method has better channel utilization under conditions of high load. However, the time delay at low load is larger. Thus, even the addition of access probability cannot overcome the inherent flaws of contention protocols.

Collision-free protocols have high delay at low load, but the channel efficiency at high load is much better then that of contention protocols. The simplest collision-free methods are TDM and FDM ([2], [16]). These allocation schemes are efficient when the number of users is small, and the traffic is continuous. Under conditions of bursty traffic, or when the number of users is large and variable, collision-free protocols are poor choices.

The limited contention protocols are based on *collision detection* capability. Namely, every collision is immediately detected by all users. Unfortunately, in some networks such as wireless networks, there is no collision detection capability: The users transmit messages through an up-link channel, and receive messages through a down-link channel. Hence, each user is only aware of collisions for messages sent by that particular user by using a time out acknowledge mechanism, but is not aware of collisions of messages from other users. These protocols are particularly impractical for satellite-based networks which have huge propagation delay and a large coverage area, since a user cannot always detect transmissions of other users, even when listening to the up-link channel (the "hidden stations" phenomenon). Listening to the down-link channel would not help because of the propagation delay. Thus, the currently available methods for sharing a limited amount of bandwidth have many drawbacks.

There is therefore a need for a protocol suitable for limited contention multiple access, which is efficient at conditions of both high load and low load. Such a protocol will be suitable for a shared channel with poor collision detection capability, and conditions of bursty traffic and a large number of users. Examples for such channels are the reservation channel in satellite-based networks, and the control channel in cellular networks.

1.1 Goals of the paper

In this paper we propose another approach to the problem of *limited contention* multiple access, that is particularly suitable for shared channels with poor collision detection capability, or long propagation delay. Previous methods of *limited contention* protocols are based on immediate collision detection capability of each user, for each time slot. As such, they are not useful for such networks. The basic idea of the proposed method is that the channel access probability depends on the load of the shared channel. The main contribution of this study is the usage of a broadcast mechanism, that already exists in many networks, to announce the users the access probability to the shared channel. Note that the proposed method is applicable also for networks in which carrier sense is reliable. In such networks, each user can estimate the load by listening to the channel. An example for this variant is a dynamic persistent CSMA, in which the access probability depends on the load. This variant of the proposed protocol can either replace a collision resolution algorithm, or enhance the performance of carrier sense protocols. However, the focus in this paper is on the other variant of the proposed strategy, to be used in networks in which carrier sense is not feasible. In this case, the network determines a *channel access probability*, which depends on the load of the shared channel, and announces it as a broadcast message to the users. Each user attempts to access the shared channel according to the announced probability. Thus, whenever the load on the shared channel is low, the users are requested to access the channel more often, while at heavily loaded time periods the users wishing to access the channel may need to wait before attempting to access the network channel. Hence, the users attempt to access the shared channel according to an announced *Dynamic Access Probability (DAP)*. Clearly, this variant of the DAP protocol requires certain requirements on the network: According to the *announced* DAP protocol, a system is provided for determining the access to a shared channel. The system comprising: (a) An access *authority* for determining an access probability to a shared channel; and (b) An access *broadcaster* for broadcasting the access probability to the users wishing to access the shared channel.

The implementation of the DAP protocol can be done on cellular and satellite-based networks. For a satellite-based network with on-board processing, the access authority is the satellite, the access broadcaster is the *existing* down-link signaling channel from the satellite and the shared channel is the up-link signaling channel for transmitting a Request To Send (RTS) data to the satellite, known as the reservation channel.

For a cellular network, the shared transmission channel includes an up-link signaling channel, the access authority is selected from the group consisting of a mobile switching center (MSC) and a base station (BS), and the access broadcaster is the base station. The announced access probability is transmitted through the existing down link control channel (e.g., DCCH in GSM systems). More preferably, for either the cellular network or the satellite network, the user attempts to transmit a Request-to-Send (RTS) message according to the access probability on the up-link signaling channel.

As opposed to other MAC protocols aimed to reduce collisions, using the DAP scheme can potentially reduce the likelihood of collisions, without causing any degradation in system performance. The advantage of the proposed method on other methods is in its load-independent channel utilization, simplicity, and robustness.

The focus in this paper is on the *announced* DAP protocol. This is since it can be implemented on networks in which collision resolution is not feasible, for example due to long delay success/failure feedback, or the existence of hidden stations. The potential benefit of using the announced DAP protocol for such networks can be very significant.

The idea of adapting the access probability to the channel load was suggested in [8] and [14]. The method suggested in those studies is known as stabilized slotted ALOHA [2]. Unfortunately, the strategies suggested in these studies are not applicable on many networks, such as satellite-based networks and cellular networks. The reason for this is that these strategies are based on immediate collision detection capability *for each time slot*, by all the users. The basic idea is that the access probability increases when there is an idle slot, and decreases when a collision occurs. In many networks this immediate collision detection capability does not exists. For example, due to the large propagation delay and the "hidden stations" phenomenon, this method cannot be used in satellite-based networks. In many other networks (e.g. cellular networks), a collision detection capability does not exist at all. Furthermore, since the computation of the desired access probability suggested in these studies is conducted per each and every time slot, these strategies are not stable, especially for bursty traffic.

For this reason, we suggest to use the statistical usage of the shared channel to estimate the optimal access probability, using a stabilization mechanism. Since an accurate load estimation is crucial to the proposed method, the main focus of this study is to reduce and evaluate the uncertainty in the load estimation. Our analysis shows that using the statistical usage of the shared channel, the optimal access probability can be well estimated for a broad class of load distribution patterns.

The concept of an *announced* dynamic access probability represents another computational approach, which is a combination of both distributed and centralized schemes. This protocol is distributed in the sense that each user makes its own decisions when to access the shared channel. On the other hand, to reduce the likelihood of collision, and improve system performance, the network, or an entity controlling the network, determines the *rate* of channel access. Hence, the *rate* of users access is centrally controlled by the system, and changes dynamically as a function of the load. The ability of the DAP strategy to combine the best features of both distributed and centralized schemes increases the efficiency of the DAP protocol at both low and high loads on the shared channel.

The channel access probability is not necessarily unique. Different access probabilities may be assigned to different users, or tasks. For example, different groups of stations may use different access probabilities, and different tasks may get different priorities.

The structure of the paper is as follows: Model and problem formulation are provided in Section 2. In Section 3 we describe the DAP protocol. Analysis is provided in Section 4, and performance analysis is provided in Section 5. Numerical results are given in Section 6, and simulation results based on real network data are given in Section 7. A summary and concluding remarks are given in Section 8.

2 Model and problem formulation

We consider a network having a shared channel. The network consists of many independent bursty stations wishing to communicate with/through a central station via the shared channel. The problem addressed in this paper is the allocation of the shared channel among many competing independent stations wishing to use it. It is assumed that new transmission attempts are generated according to a Poisson distribution. To a good approximation, the number of users may be considered as infinite, for traffic considerations. It is assumed that the time is slotted, and transmissions occur only at the beginning of a time slot. All packets are of the same length, and each time slot can accommodate the transmission of exactly one packet.

Examples for such network are: Satellite-based networks with on-board processing, in which many ground stations communicate through a satellite, and cellular networks, in which at each cell many mobile users communicate with a base station. It is assumed that there is a broadcast mechanism from the central station to all the stations, that all the stations listen to. Hence, a message transmitted through the broadcast channel is received by all the stations.

3 The Announced Dynamic Access Probability protocol (DAP)

Below we describe the announced dynamic access probability protocol. The network computes an access probability, which depends on the load, and announces it, as a broadcast message, to the users. Each user who wishes to use the shared channel, attempts to access this channel with the announced probability.

The access probability is determined by the load: It increases at low load periods and decreases at high load periods. By listening to the shared channel, and counting (or estimating) the number of transmission attempts per time unit, the system attempts to optimize the access probability to achieve an optimal channel efficiency.

The shared channel access algorithm is as follows:

- Network manager:
  Initial procedure:
  At start, set the access probability to one: $p_{access} = 1$.
  Repeat:
  1. Estimate the average number of transmission attempts per time unit, say r, by listening to the channel.
  2. Set the user access probability p_{access}, based on the value of r, and the last used value of p_{access}, and announce it to all users, as a broadcast message.
- User access:
  Conceptually, there are two variants:
  - A) Accessing a signaling channel (e.g., the control channel in cellular networks, the reservation channel in satellite-based networks). To access the data channel, a collision-free protocol, such as TDMA or FDMA ([2], [16]), should be used:
    1. Get access probability, say p_{access}.
    2. Attempt to send a Request-To-Send (RTS) message, with probability p_{access}.
    3. Wait for acknowledge (Clear-To-Send [CTS] message).
    4. When CTS is received, send data, following the instructions specified in the received CTS message.
  - B) Direct access to the data channel, in case there is no signaling channel:

1. Get access probability, say p_{access}.
2. Attempt to access the shared channel with probability p_{access}.

4 Analysis

The implementation of the DAP protocol requires, among other things, two main tasks: An accurate estimation of the load on the channel, and the computation of the optimal access probability p_{access} for that load. These issues are discussed in this section.

4.1 Load Estimation

Since the access probability depends on the load, an accurate estimation of the load is crucial for computing the optimal access probability. The load average estimation is based on the statistical usage of the channel.

We distinguish between two types of time intervals: The time interval used to estimate the load, denoted as the *estimation interval*, and the time interval in which the access probability computed during the *estimation interval* is applied. We refer to this interval as the *announced interval*, since the access probability that was estimated during the *estimation interval* is announced during the *announced interval*. The length of the *announced interval* does not have to be necessarily equal to the length of the *estimation interval*. However, to simplify the analysis, we consider the case in which both intervals are equal.

4.1.1 The length of the estimation interval

Let τ be the length, in terms of number of time slots, of the *estimation interval*. The uncertainty in the load estimation depends on two major factors: The statistical error of estimating the average load using a finite time interval, and the change in the average load over time. The dynamic behavior of the average transmission attempt rate implies that an important part of the load estimation is the choice of the interval length for measuring the load, denoted by τ. The desired value of τ is chosen such that it is sufficiently large to be statistically reliable, and sufficiently small such that the load change from the *estimation interval* to the *announced interval* is negligible.

Let $\lambda(t)$ be the rate of transmission demands at time t (i.e. the rate of users wishing to transmit at time t is $\lambda(t)$). The *expected* number of transmissions per time slot (both successful and unsuccessful) during the time interval $[t, t+\tau]$ is therefore given by:

$$\theta_t = \frac{1}{\tau}\int_t^{t+\tau} \lambda(s)ds. \tag{1}$$

Let $\tilde{\theta_t}$ be the average number of transmissions per time slot, as measured along the time interval $[t, t+\tau]$. Given that $\tau = n$ time slots (i.e. n samples), we get that $\tilde{\theta_t}$ is the arithmetic mean of the sum of n independent and identically distributed random variables. It is shown in [7] that Chebyshev's inequality implies that for any $\epsilon > 0$ we get that:

$$Pr[|\tilde{\theta_t} - \theta_t| \geq \epsilon] \leq \frac{Var(X)}{n\epsilon^2}. \tag{2}$$

Where X is a random variable denoting the number of arrivals during a single time slot, and $Var(X)$ is the variance of X. Note that if the arrival process is a Poisson process, one can easily derive $Var(X)$. Equation 2 implies that $\tilde{\theta_t}$, which is the arithmetic mean of the sum of n independent and identically distributed random variables will approach its expected value θ_t as n increases. Unfortunately, the random variable X is time dependent. Hence, n should be sufficiently small, such that we can assume that there is no significant change in the expected transmission attempts rate during n time slots. Equation 2 holds for any load pattern, assuming that the arrivals during the time interval $[t, t+\tau]$ are uncorrelated. This assumption can be easily justified by using the rule that each station must wait at least n time slots before re-transmitting.

Our goal is to find a proper value of τ such that $\tilde{\theta_t}$ should accurately predict the expected arrival rate at the *announced interval*. For example, consider a system in which a time slot equals 10 milliseconds, and the average load increases by 10% in a minute. Choosing $\tau = 10$ seconds, we get that $\tau = 1000$ time slots, and the expected load increment during $\tau = 10$ seconds is $|\theta_{t+1} - \theta_t| \approx 1.5\%$, which is negligible for most practical cases. The time interval length used for load estimation, as well as the time interval length in which p_{access} is applied, should both be long enough in order to be stable and noise-insensitive, and at the same time to be short enough in order to have, for a good approximation, a constant rate of new transmission attempts, and negligible load fluctuations. Since our goal is to compute the optimal channel access probability p_{access} to be used during the *announced interval*, we must determine the desired length of the *announced interval*. Since p_{access} is to be used during at least τ time slots, the value of τ should be sufficiently small to accommodate one fixed value of p_{access} and insure that the amount of fluctuations in $\lambda(t)$ during the *announced interval* is minimal.

In our analysis the length of the estimation interval is equal to that of the announced interval. This however is not mandatory. For example, past history of the network may be used for extrapolation. A detailed discussion of the problem analyzed above can be found in [4].

Under the condition that new transmission attempts are generated according to a Poisson distribution, it is shown in [7] that, for sufficiently long τ (e.g. $\tau \approx 100$ time slots is sufficiently large), $\tilde{\theta}_t$ is normally distributed around θ_t. That implies that estimating θ_t by measuring $\tilde{\theta}_t$, the standard deviation of $\tilde{\theta}_t$ from θ_t is given simply by $\sqrt{\frac{\theta_t}{\tau}}$:

$$E[(\tilde{\theta}_t - \theta_t)^2] = \frac{\theta_t}{\tau}. \tag{3}$$

Figure 1 depicts the relative load estimation error as a function of the transmissions rate, given that new transmission attempts are generated according to a Poisson distribution. The relative load estimation error, defined as the standard deviation of $\tilde{\theta}_t$ from θ_t divided by θ_t, is depicted as a function of θ_t, for $\tau = 1000$ time slots. It can be seen that the relative error is not larger than 10%, for $\theta_t \geq 0.1$, and decreases with the transmissions rate θ_t. Given that the load estimation relative error is 10%, the expected channel utilization under the DAP protocol is bounded from below by $min\{1.1e^{-1.1}, 0.9e^{-0.9}\} = 0.35$. The relative drop in the channel utilization, in comparison to the optimal achievable value, is therefore $1 - \frac{0.35}{e^{-1}} = 1 - 0.953 < 0.05$. Hence, Figure 1 implies that, given a Poisson load distribution, the length of the *estimation interval* should range from one second to 10 seconds, depending on the time slot size. The condition that there is no significant change in the load during this time interval holds for most wireless networks, and is violated very rarely in practice.

4.2 Channel utilization

Let n_{users} be the number of users wishing to access the shared channel. Each user attempts to access the channel with probability p_{access}. The channel utilization U is the probability that exactly one user transmits:

$$U = \binom{n_{users}}{1} p_{access}(1 - p_{access})^{n_{users}-1} =$$

$$n_{users} p_{access}(1 - p_{access})^{n_{users}-1}. \tag{4}$$

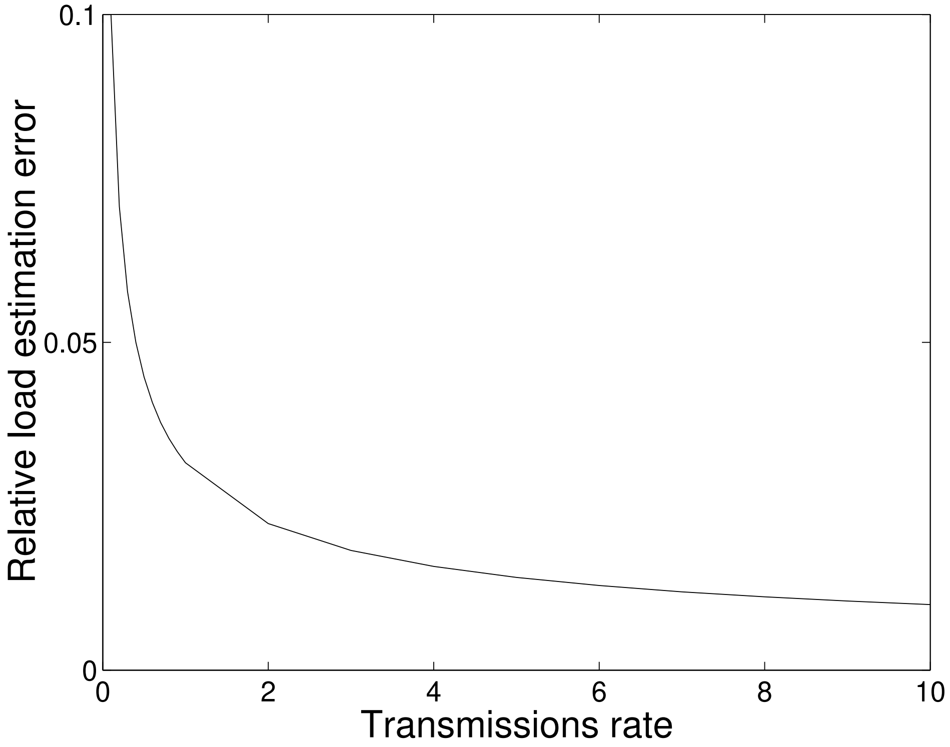

Figure 1: Relative load estimation error as a function of the transmissions rate, when new transmission attempts are generated according to a Poisson distribution.

Differentiating U with respect to p_{access} and set the result to zero, we get that the maximal value of U is achieved for $p_{access} = \frac{1}{n_{users}}$. Substitute $p_{access} = \frac{1}{n_{users}}$ in Equation 4 we get that:

$$U_{max} = (1 - \frac{1}{n_{users}})^{n_{users}-1}, \tag{5}$$

note that $U_{max} \to \frac{1}{e}$ when $n_{users} \to \infty$. Hence, using $p_{access} = \frac{1}{n_{users}}$ for a highly loaded system (n_{users} is very large), the channel utilization under DAP equals to the optimal channel utilization of slotted ALOHA with Poisson arrivals, and an infinite number of users, as expected.

Clearly, the expected number of transmissions (both successful and unsuccessful) per time slot r depends on the number of users wishing to transmit according to:

$$r = n_{users} p_{access}. \tag{6}$$

4.3 Channel feedback

The method of estimating the load on the channel depends on the available channel feedback. We distinguish between three situations:

- A *binary feedback*: the feedback from the channel distinguishes between an empty (idle) slot, and a non-empty slot (one or more transmission attempts).
- A $0/1/C$ feedback, in which we can distinguish between an empty slot ('0'), a successful (single) transmission ('1'), and a collision ('C')- when at least two transmissions occur at the same time slot.
- No channel feedback: the only feedback is when a packet arrives to the central station.

Assuming that transmission attempts, both new and backlogged, are generated according to a Poisson process with mean r, the probability of k transmission attempts per time slot is given by the Poisson distribution:

$$Pr[K \; transmission \; attempts] = \frac{r^k e^{-r}}{k!}. \tag{7}$$

Where the term "transmission attempts" is the number of transmissions, both successful and unsuccessful. Hence, the probability of an empty slot is given by:

$$Pr[0 \; transmission \; attempts] = e^{-r} \tag{8}$$

Equation 8 yields a direct estimation of r by counting the average number of empty slots. The channel utilization is the probability of exactly one transmission attempt per time slot. Using Equation 7, it is given by:

$$U = re^{-r}, \tag{9}$$

The average number of transmission attempts per time slot r can be easily estimated using Equations 7- 9. Hence, for a $0/1/C$ feedback, in addition to using Equation 8, we use Equation 9 to get that the probability of a successful transmission (exactly one transmission attempt during a single time slot) is given by re^{-r}. The arithmetic average of both estimations yields a reasonable approximation of r. The probability of a collision is given therefore by:

$$Pr[collision] = 1 - e^{-r} - re^{-r}. \tag{10}$$

Under optimal channel efficiency, we substitute $r = 1$ to get that the target collision probability is $1 - \frac{2}{e} \approx 26\%$. The arithmetic average of these estimations (empty/successful/collision) yields another reasonable approximation of r.

The most difficult situation is when there is no channel feedback. For this situation we assume that each packet has a number which identifies the packet's transmission attempt. The first transmission is number one, the first re-transmission is number two etc... Since the central station receives all packets from the other stations, it can estimate r by using the fact that the expected number of transmission attempts is given by e^r [16]. Hence, the average of $\ln(num)$, where num is the number of transmission attempts of a packet, yields a reasonable approximation of r.

4.4 Channel access probability computation

The computation of the optimal access probability p_{access} is based on the statistical usage of the shared channel. Let λ be the expected rate of users wishing to transmit per time slot (both new arrivals and backlogged). Given a constant access probability p_{access}, the dependency of the expected rate of transmissions per time slot (both successful and unsuccessful) denoted by r, on λ is given by:

$$r = p_{access}\lambda. \tag{11}$$

The average rate r of transmissions per time slot, along time interval of length τ time slots, where τ is the parameter discussed in Section 4.1, is used to determine the access probability for the next τ time slots. The time interval τ is considered sufficiently small such that the average rate r does not change significantly along 2τ time slots. From now on the time interval $[t, t+\tau]$ is denoted by t, and the average rate of transmissions per time slot, both successful and unsuccessful, along the time interval $[t, t+\tau]$ is denoted by $r(t)$.

Let r_{opt} be the optimal rate of transmissions, under which the channel efficiency is maximized. The computation of r_{opt} is conducted assuming that new transmission demands, both new and backlogged, are generated according to a Poisson process, with mean λ transmission demands per time slot, and given that each user uses the access probability p_{access}. Under these conditions, we get that the process consists of *all* transmission attempts, both new and backlogged, is a Poisson process (see, e.g.[2]), with mean r transmissions per time slot, both successful and unsuccessful. The value of r is given in Equation 11 as a function of λ and p_{access}. Differentiating Equation 9 with respect to r, set the result to zero, and solving for r, we get that the maximal channel efficiency is at $r_{opt} = 1$. Substitute $r = \lambda p_{access}$ in Equation 9, and using the fact that p_{access} is constrained to be $0 \leq p_{access} \leq 1$, we get that the value of p_{access} that maximizes the channel efficiency is given by:

$$p^*_{access} = min\{\frac{1}{\lambda}, 1\}. \tag{12}$$

Hence, given an average transmission rate $r(t)$ at time interval t, the new access probability at time interval $(t+1)$ is given by:

$$p_{access}(t+1) = min\{\frac{p_{access}(t)}{r(t)}, 1\}. \tag{13}$$

Hence, if $r(t) > 1$, then $p_{access}(t+1) < p_{access}(t)$. If $r(t) \leq 1$, then $p_{access}(t+1) \geq p_{access}(t)$. Substitute $r = 1$ in Equation 9, we get that the optimal channel utilization is given by:

$$U = e^{-1} \approx 37\%. \tag{14}$$

4.4.1 The stabilization mechanism

Our goal next is to keep the announced access probability p_{access} as stable as possible, without reducing the channel utilization. In order to achieve this goal, the transmissions rate $r(t)$ must be kept within the range $r_{opt} - \epsilon \leq r \leq r_{opt} + \epsilon$, where ϵ is a small pre-defined parameter that depends on the error of the load estimation. The parameter ϵ is chosen such that it satisfies:

$$max\{|U(r_{opt}) - U(r_{opt} + \epsilon)|, |U(r_{opt}) - U(r_{opt} - \epsilon)|\} \leq \delta, \tag{15}$$

where $r_{opt} = 1$ and δ is a pre-defined parameter that reflects the upper bound on the channel utilization degradation, that we are willing to accept as a result from the error of the load estimation. Equation 15 implies that the channel utilization at $r_{opt} \pm \epsilon$ may differ from the channel utilization at r_{opt} by at most δ. For example, if $\delta = 0.05$ then $\epsilon = 0.3$, since $0.7e^{-0.7} = 0.347$, and $1.3e^{-1.3} = 0.354$, as compared to the optimal channel utilization $e^{-1} = 0.367$. **That implies that even if the error in the load estimation is up to 30%, the DAP performance, in terms of the channel utilization, may degrade by at most 5%.** In other words - the channel utilization under the DAP protocol is not sensitive to the accuracy of the load prediction.

The access probability p_{access} is therefore computed as follows: If $r_{opt} - \epsilon \leq r(t) \leq r_{opt} + \epsilon$, then $p_{access}(t)$ remains unchanged. Otherwise, Equation 13 is used to compute the new access probability whenever $|r(t) - r_{opt}| > \epsilon$.

The proposed stabilization mechanism can be used to stabilize the access probability around its optimal value. As a consequence, the access probability decreases with the load: Whenever the transmission rate is above a pre-defined threshold, it decreases. Whenever the transmission rate drops below another pre-defined threshold, it increases and approaches to 1. A Hysteresis curve may be used to stabilize the access probability around its optimal value $p_{access} = \frac{1}{r}$. Hence, *using the DAP protocol guarantees, at high load periods, a channel efficiency of approximately* $\frac{1}{e}$*, that depends on the load variability.*

5 Performance Analysis

Below we analyze the performance of the DAP protocol in comparison to slotted ALOHA. The ALOHA protocol and its many variants is widely used in real systems, especially in wireless networks. The analysis is conducted on a discrete time system, where the time is slotted. The focus of the performance analysis of the DAP protocol is on two basic parameters: The *channel utilization*, which reflects the protocol efficiency, and the time interval required for successful transmission, which reflects the waiting time for message delivery. We show that the channel utilization is relatively high at high load, while the delay at low load periods is minimal.

5.1 Channel utilization

The performance of slotted ALOHA has been widely discussed and analyzed. Following the analysis in [16], and using Equation 5, we get that the channel utilization under DAP equals to the optimal channel utilization of slotted ALOHA with Poisson arrivals, and an infinite number of users, as expected. However, while the performance of slotted ALOHA decreases exponentially with the number of users attempting to access the channel, the channel utilization of DAP is, for a good approximation, load-independent.

Figure 2 depicts the channel utilization U of DAP using Equation 5, in comparison to slotted ALOHA and static p-persistent slotted ALOHA, for $p = 0.1$ and $p = 0.2$, as a function of λ - the expected rate of users wishing to transmit. Under conditions of light load the performance of static p-persistent slotted ALOHA is very poor, while under conditions of high load the performance of slotted ALOHA is very poor. As the load increases, the channel utilization of the DAP protocol becomes load-insensitive, while the channel utilization of all other protocols depicted in Figure 2 approaches zero as the load increases.

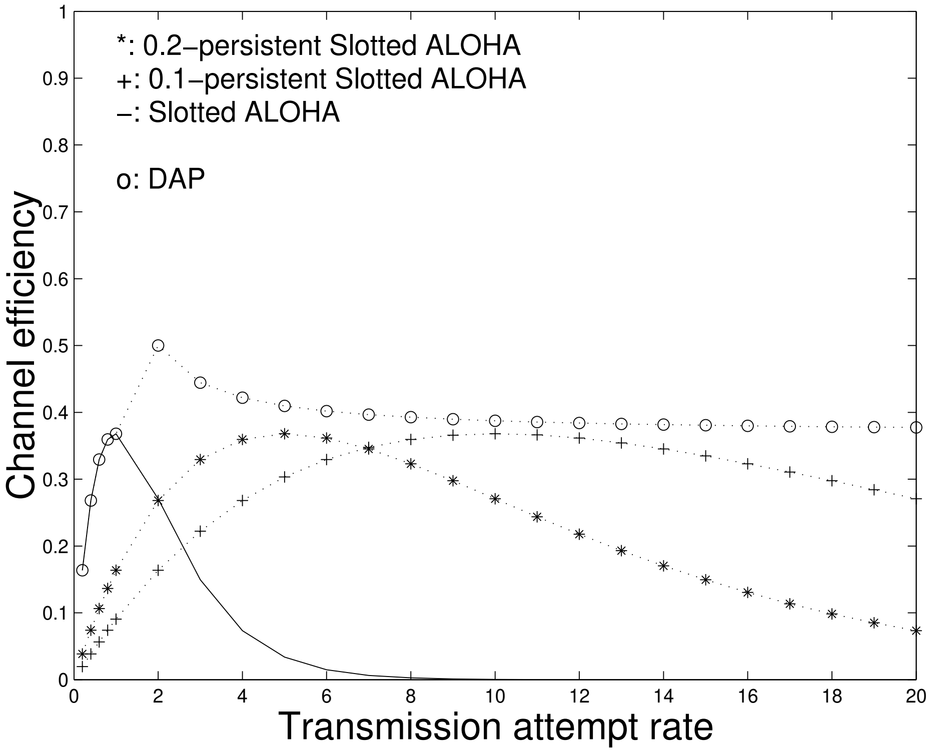

Figure 2: Channel efficiency of the DAP protocol, in comparison to slotted ALOHA and p-persistent slotted ALOHA

It is interesting to note that, for small number of users wishing to transmit per time slot, the channel utilization U of DAP is greater than $\frac{1}{e}$, and decreases asymptotically with λ.

5.2 Expected time for successful transmission

The channel access probability p_{access} affects the time interval required for a successful transmission in two contradicting ways: On one hand, the expected time interval between two successive transmissions decreases with p_{access}, indicating that under conditions of low traffic p_{access} should approach 1. On the other hand, as we show below, the expected number of transmission attempts required for a successful arrival increases exponentially with p_{access}.

Following the analysis in [16], and using Equation 8, the probability to find an empty slot is given by e^{-r}. Consequently, the probability that a successful transmission requires exactly k attempts ($k-1$ collisions followed by one success) is given by $e^{-r}(1-e^{-r})^{k-1}$ (assuming that the time interval required for a successful transmission is sufficiently small, such that r can be considered as a constant). Hence, the expected number of attempts required for a successful transmission is given by:

$$E[k] = \sum_{j=1}^{\infty} j e^{-r}(1-e^{-r})^{j-1} = e^{r}. \tag{16}$$

Equation 16 implies that a small increases in r can drastically reduce the channel efficiency of slotted ALOHA.

Since r is assumed constant along the time interval required for a successful transmission, $p_{access}(t)$ is also constant, and denoted simply by p_{access}. Using a similar analysis to that leading to Equation 16, the expected number of time slots required for one transmission attempt is given by:

$$E[d] = \sum_{j=1}^{\infty} j p_{access}(1-p_{access})^{j-1} = \frac{1}{p_{access}}. \tag{17}$$

Using equations 6, 11, 16, and 17, the expected number of time slots required for a successful transmission is given by:

$$D = \frac{e^{r}}{p_{access}} = \frac{e^{\lambda p_{access}}}{p_{access}}. \tag{18}$$

Differentiating D with respect to p_{access} and set the result to zero, we get that the minimal value of D is achieved for $p_{access} = \frac{1}{\lambda}$. For $p_{access} < \frac{1}{\lambda}$ we get that $\frac{\partial D}{\partial p_{access}} < 0$, implying that at conditions of light load ($r = \lambda p_{access} < 1$) the waiting time decreases with p_{access}, since the time interval between successive transmission attempts decreases with p_{access}, without sufficiently increasing the likelihood of collisions. On the other hand, for $p_{access} > \frac{1}{\lambda}$, we get that $\frac{\partial D}{\partial p_{access}} > 0$, implying that at conditions of high load ($r = \lambda p_{access} > 1$), the expected time for a successful transmission increases with p_{access}. This is since the time interval between successive transmission attempts indeed decreases linearly with p_{access}, but the number of collisions increases exponentially with

p_{access}. Hence, it follows from Equation 4 and Equation 18 that *the same value of p_{access} that maximizes the channel utilization, also minimizes the time delay.* This result is in agreement with the results obtained in [6] and [12]. Substitute $p_{access} = \frac{1}{\lambda}$ in Equation 18, we get that the minimum delay is given by:

$$D = \lambda e^{\lambda p_{access}} = e\lambda. \tag{19}$$

Equation 19 implies that the time delay linearly increases with the number of users wishing to access the channel. Since the channel utilization approaches asymptotically to $\frac{1}{e}$ (Equation 5), when the rate of users entering the system is greater than $\frac{1}{e}$, we get that λ and the time delay become both unbounded.

Figure 3 depicts the natural logarithm of the expected time required for a successful transmission of DAP, in comparison to slotted ALOHA and static p-persistent slotted ALOHA, for $p = 0.1$ and $p = 0.2$, as a function of λ - the average number of users wishing to transmit per time slot. It clearly shows that the DAP protocol outperforms the slotted ALOHA protocol and its static p-persistent variants. Using p-persistent slotted ALOHA under conditions of light load, the expected time delay of successful transmission is very large. Moreover, as the load increases, under very high load conditions the time delay under p-persistent slotted ALOHA is larger than under the DAP protocol. For example, for expected transmission demands rate greater than 8 transmission demands per time slot, the expected delay under 0.2-persistent slotted ALOHA is larger than that under the DAP protocol.

5.3 Performance analysis under bursty traffic condition

In many systems, data traffic is extremely bursty. Maximum load to average load ratios of 100:1 are common. Most of the traffic generated during relatively short time periods, while at most of the time the system is lightly loaded. Due to the exponential dependency of the channel utilization under slotted ALOHA on the load, an ALOHA system is unstable for large users population ([6], [11], [12]), even if its average load is smaller than $\frac{1}{e}$.

On the other hand, the DAP protocol, due to its capability to adjust the access probability to the load, can lead to a *load smoothing.* Consequently, under conditions of bursty traffic, the channel utilization of DAP protocol is significantly larger than that of the ALOHA family. This superiority over slotted ALOHA is demonstrated below for the following traffic pattern: We consider a bursty

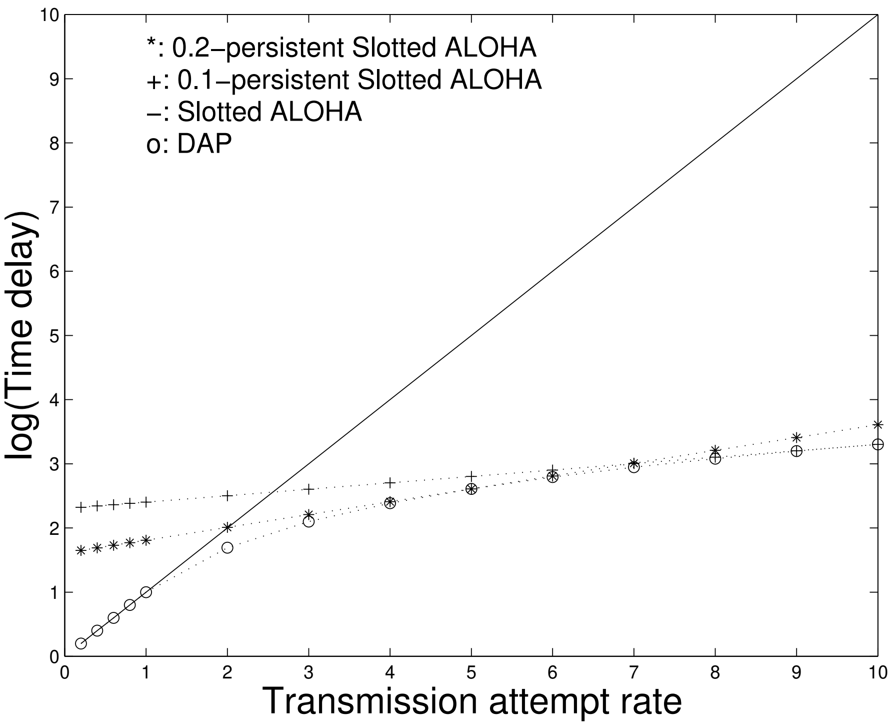

Figure 3: The expected time required for a successful transmission of the DAP protocol, in comparison to slotted ALOHA, on a logarithmic scale, as a function of the average number of users wishing to transmit per time slot.

load pattern consisting of a *heavily loaded interval*, followed by a *lightly loaded interval.* It is assumed that the average load is less than $\frac{1}{e}$.

Let τ_H and τ_L denote the length of the *heavily loaded interval* and the *lightly loaded interval*, respectively. Let λ_H and λ_L denote the average rate of new arrivals per time slot at the *heavily loaded interval* and the *lightly loaded interval*, respectively. We further assume that at the system starting point the system is empty. Figure 4 depicts the load pattern described above.

Our concern is in the range of parameters (λ_H,λ_L, τ_H,τ_L) for which the number of new arrivals during the *heavily loaded interval* and the *lightly loaded interval*, is not greater than the number of packets transmitted successfully during those time intervals.

Using the DAP protocol, the channel utilization is $\frac{1}{e}$ at the *heavily loaded interval.* The average

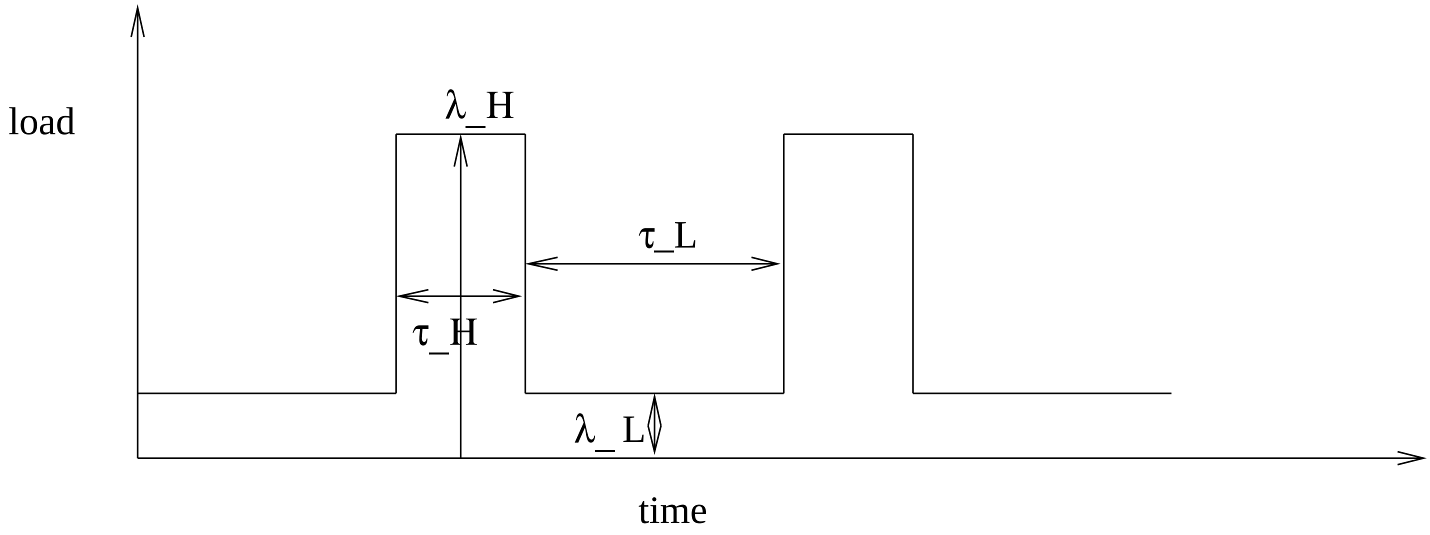

Figure 4: A bursty load pattern. The arrival rate (load) is depicted as a function of time.

load ρ is given by:

$$\rho = \frac{\tau_H \lambda_H + \tau_L \lambda_L}{\tau_H + \tau_L}. \tag{20}$$

The stability condition for the DAP protocol is:

$$\rho < \frac{1}{e}. \tag{21}$$

Equations 20- 21 imply that, under the condition that the system should be empty at the end of the *lightly loaded interval*, the parameters λ_H, λ_L, τ_H, τ_L, must satisfy:

$$\tau_H(\lambda_H - \frac{1}{e}) < \tau_L(\frac{1}{e} - \lambda_L). \tag{22}$$

Remark 5.1 *Equation 22 implies that given a burst of length τ_H and rate λ_H, in order to keep a DAP system stable, then for time interval of length at least τ_L, the average arrival rate to the system must be at most λ_L, where all these parameters must satisfy Equation 22.*

On the other hand, due to its exponential behavior, an ALOHA system cannot recover from sufficiently large burst. Hence, it is unstable for large users population ([6], [11], [12]).

6 Numerical results

To demonstrate the performance of the DAP protocol under various load distribution conditions we examine several numerical examples. Figure 5 depicts the channel utilization under DAP as a function of the relative load estimation error. It is shown that even if the load estimation error is 50%, the channel utilization degradation is at most 10%.

To study the behavior of a DAP system under a bursty traffic condition, we consider a system having n backlogged packets. Our interest is to find the maximum new arrivals rate that satisfies the condition that the rate of packets transmitted successfully is greater or equal to the rate of new arrivals. We consider this rate as the maximum "permitted" rate. Figure 6 depicts the maximum new arrivals rate that satisfies this condition for a DAP system, as a function of the number of backlogged packets, in comparison to slotted ALOHA. Since the departure rate under DAP equals e^{-1}, the maximum "permitted" rate is independent of the number of backlogged packets, and equals $\frac{1}{e}$. However, the expected time for successful transmission does increase with the number of backlogged packets, as it is shown in Figure 3. On the other hand, under slotted ALOHA the "permitted" rate is exponentially dependent on the number of backlogged packets, and equals ne^{-n}, where n is the number of backlogged packets. These behaviors are clearly shown in Figure 6. Whenever the number of backlogged packets in ALOHA system is greater than 8 packets, the departure rate is actually zero, which explains the instability of slotted ALOHA, as shown in ([6], [11], [12]).

To further study the DAP behavior under a bursty traffic condition, in comparison to slotted ALOHA, we consider the minimal time interval required to empty a heavily loaded system. Given that there are n backlogged packets, we define the *minimal recovery interval* as the time interval required to bring the system to an empty state (i.e. with no backlogged packets), under the condition of zero new arrival rate. Figure 7 depicts the *minimal recovery interval* of the DAP protocol, in comparison to slotted ALOHA, as a function of the number of backlogged packets. The *minimal recovery interval* of the DAP protocol increases linearly with the number of backlogged packets, due to it constant throughput. On the other hand, using slotted ALOHA, this time interval increases exponentially with the number of backlogged packets.

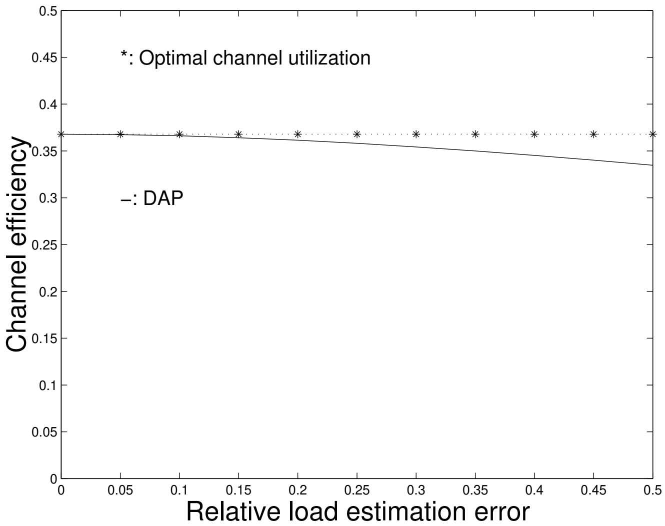

Figure 5: Channel utilization under DAP as a function of the relative load estimation error.

7 Simulation results using real network data

To demonstrate the performance of the DAP protocol under realistic conditions, the DAP protocol was applied on a real data of Internet traffic. The data is taken from a Web Cache Proxy log file (taken from the depository of the IRCACHE project, see [18]). A Web proxy log file records all the Web requests made to Web pages by a certain population of Web users (which direct their requests to the Web Proxy instead of directing them directly to the Web). This stream of requests, thus represents a typical stream of requests coming from a population of Web users (e.g the set of users of a certain Internet Service Provider), and thus reflects a typical stream that would be directed into a satellite system servicing Web applications. Note that since Web requests (as opposed to Web pages) typically consist of a single packet, a simulation that is based on a sequence of packets whose time stamps are taken from the Web log trace file, should reflect a real transmission request stream experienced in a satellite system. The users access requests, stored in the proxy log file, represent new arrivals (i.e. new transmission wishes). A feedback mechanism was used, such that each collided packet was re-transmitted as a backlogged packet. Figure 8 depicts the channel utilization of the DAP protocol, in comparison to p-persistent slotted ALOHA: 0.1-persistent ALOHA, 0.2-persistent slotted ALOHA, and 0.5-persistent slotted ALOHA. The protocols were applied on a 2500 seconds long time interval, in which the average rate of new arrivals was 0.29 new arrivals per time slot. Each point at Figure 8 depicts the average channel utilization over a 2.5 seconds long

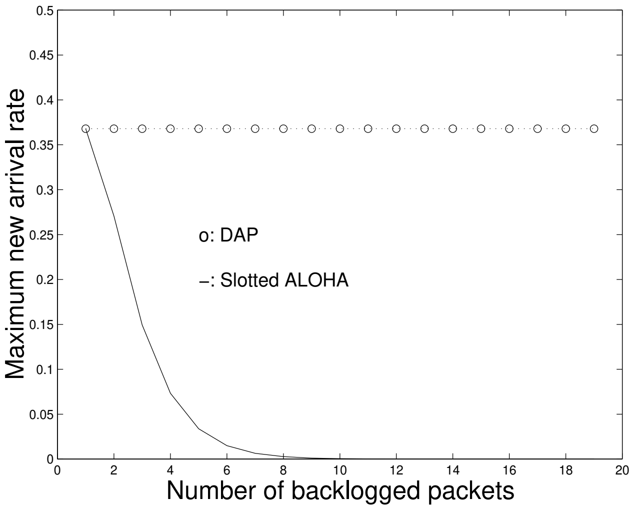

Figure 6: Maximum "permitted" new arrival rate, as a function of the number of backlogged packets.

time interval. As it is shown in Figure 8, the DAP protocol could overcome this traffic, and the average of its channel utilization was 0.29 - implying that all packets were successfully transmitted. On the other hand, all variants of p-persistent ALOHA could not handle the rate of new arrivals, and after a relatively short time period of $3-4$ minutes, their utilization was nearly zero, since all packets were backlogged. This is in agreement with the observation that the expected channel utilization of slotted ALOHA and its many variants in real networks is, for a good approximation, about 0.2 (in comparison to the optimal value of 0.37). Most importantly, as it is shown in Figure 8, even when the DAP protocol did not predict accurately the rate of new arrivals, it has soon recovered. Consequently, the channel utilization, that has dropped during some time intervals to 0.15, increased soon after that to 0.37. Hence, the DAP protocol can handle real traffic (not necessarily Poisson) with an expected rate of at least 0.29 new arrivals per time slot.

To further understand the behavior of real (Internet) traffic, and the performance of the DAP protocol, the traffic variability was studied. Figure 9 depicts the optimal access probability that should be used during the same time interval used in Figure 8. Each point at Figure 9 depicts the average optimal access probability over a 2.5 seconds long time interval. The optimal access probability was computed using Equation 12, and the actual number of transmission attempts (both new arrivals and backlogged packets), assuming an exact load prediction for each and every

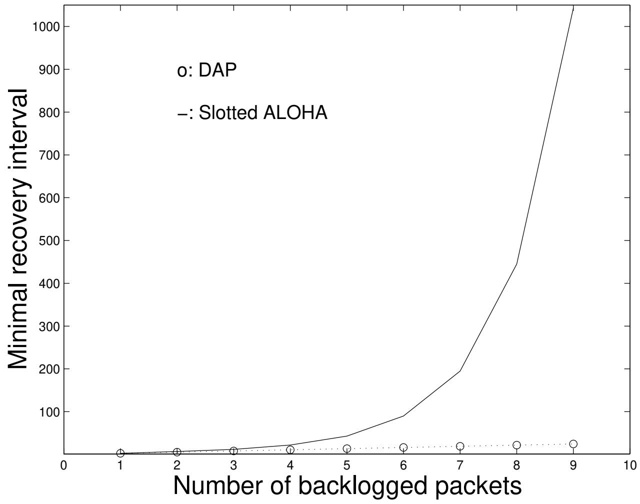

Figure 7: The minimal recovery interval, as a function of the number of backlogged packets.

time slot. As it is shown in Figure 9, the optimal access probability varies from 0.05 to 1 during relatively short time period of about 42 minutes. This variability, which expresses the variability in the transmission attempt rate, can explain the superiority of a dynamic access probability (DAP) protocol over a static p-persistent protocol.

To further study the DAP performance under realistic conditions, the algorithm was applied on a real data of a satellite-based network, provided by Gilat satellite networks Ltd. Since the expected rate of new arrivals in this network was very high - about 0.4 new packets per time slot (the network has more than one data channel), and in order to study the maximum achievable throughput boundary of the DAP protocol, under real traffic conditions, no feedback mechanism was applied, and colliding packets were simply dropped. Figure 10 depicts the channel utilization under DAP in comparison to slotted ALOHA, as a function of time, measured in time slots. It demonstrates very clearly the superiority of the DAP protocol over slotted ALOHA, under conditions of real traffic. The results show that the expected channel utilization under DAP is about 0.32. Using slotted ALOHA for the same traffic data the expected channel utilization is 0.21. It is interesting to note that the achieved channel utilization under DAP is relatively stable, and ranges from 0.31 to 0.33.

Figure 11 depicts, for the same data used for Figure 10, the expected time required for a successful transmission of the DAP protocol, in terms of time slots, in comparison to p-persistent slotted ALOHA, using Equation 18. Clearly, under conditions of real traffic, the time delay of p-persistent slotted ALOHA is significantly larger than that of the DAP protocol.

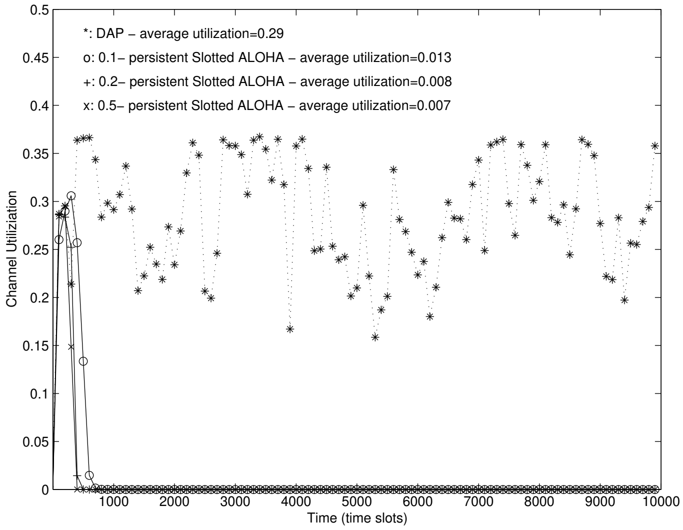

Figure 8: Channel utilization, under conditions of real traffic.

8 Summary and concluding remarks

This paper proposed a dynamic multiple access protocol, which is efficient yet reliable. The basic idea is to use an auxiliary control to stabilize a contention protocol, such as the slotted ALOHA or CSMA, around its optimal working point. As a consequence, the channel utilization is load-independent. This is achieved without increasing the time delay at low load periods. We show that using the statistical usage of the shared channel, the optimal access probability can be well estimated for a broad class of load distribution patterns. As the traffic is more bursty, the DAP strategy offers further performance improvement, in comparison to its equivalent load insensitive protocol (e.g. slotted ALOHA). Hence, the DAP protocol is especially suitable for networks with unreliable channel feedback, in which the number of users is very large and the traffic is bursty. Examples for such networks are satellite-based networks with on-board processing and cellular networks.

Acknowledgment

This work was supported by the ISIS R&D Consortium administered by the Chief Scientist of the Israeli Ministry of Industry & Trade. Partial results of this study were reported at INFOCOM 2001.

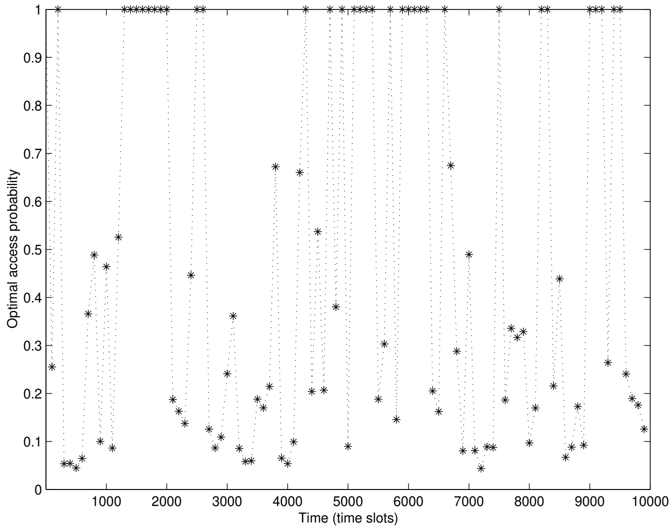

Figure 9: Optimal channel access probability, under conditions of real traffic.

References

[1] Abramson, N., "Development of the ALOHANET", IEEE transaction on information theory, vol. IT-31, pp. 119-123, 1985.

[2] Bertsekas D. and Gallager R., Data Networks, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1992.

[3] Bharghavan V., Demers A.,Shenker S., and Zhang L., MACAW: A Media Access Protocol for Wireless LANs, Proc. SIGCOMM'94 conf., ACM, pp. 212-225, 1994.

[4] Bikel P.J. and K.A. Doksum, Mathematical statistics- basic ideas and selected topics, Holden-day inc., 1977.

[5] Capetanakis J.I., Tree Algorithms for Packet Broadcast Channels, IEEE trans. on information theory, vol. IT-25, pp. 505-515, 1979.

[6] Fayolle G., Gelenbe E., and Labetoulle J., Stability and optimal control of the packet switching broadcast channel, journal of the association for computing machinery, Vol. 24, No. 3, 1977, pp. 375-386.

[7] Feller W., An Introduction to Probability Theory and Its Applications, Vol. 1, John Wiley, 1968.

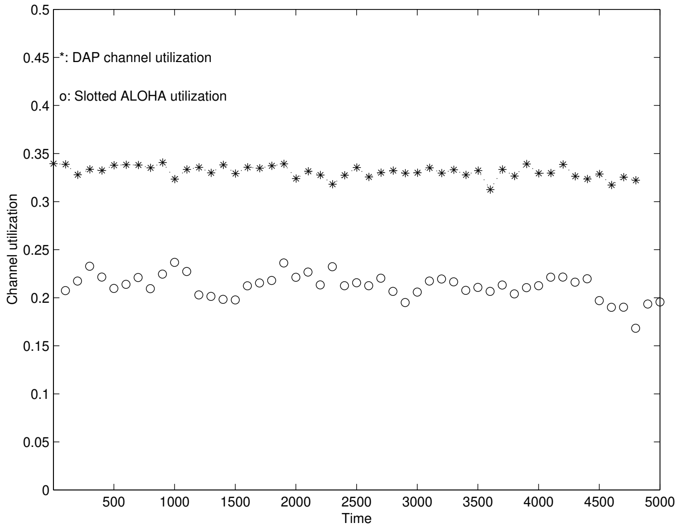

Figure 10: The channel efficiency of the DAP protocol, as a function of time, in comparison to slotted ALOHA, under conditions of real traffic.

[8] Hajek B. and Loon T., Decentralized dynamic control of a multiaccess broadcast channel, IEEE transactions on automatic control, Vol. 27, No.3, 1982, pp. 559-569.

[9] Karn P., MACA - A new channel Access Protocol for Packet Radio , ARRL/CRRL Amateur Radio Ninth computer networking conf., pp. 134-140, 1990.

[10] Kleinrock L. and Tobagi F., Random Access Techniques for Data Transmission over Packet-Switched Radio Channels, Proc. Nat. computer conf., pp. 187-201, 1975.

[11] Kleinrock L. and S. Lam, Packet switching in a multiaccess broadcast channel: Dynamic control procedures, IEEE transactions on communications, Vol. 23, 1975, pp. 891-p04.

[12] Kleinrock L., Queueing Systems, Vol. 1,2, Wiley, 1976.

[13] Nanda S., Balachandran K., and Kumar S., Adaptation Techniques in Wireless Packet Data Services, IEEE communication magazine, pp. 54-64, January 2000.

[14] Rivest R.L., Network control by Bayesian broadcast, MIT report No. 285, 1985.

[15] Roberts L., "Extensions of packet communication technology to a hand held personal terminal", Proc. SJCC, pp. 295-298, 1972.

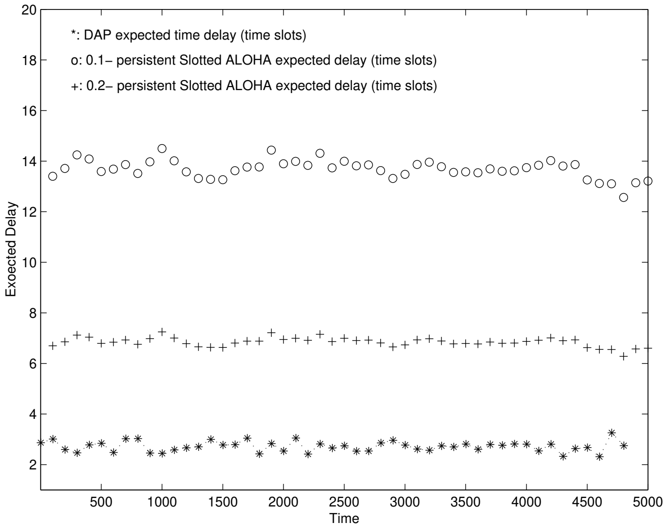

Figure 11: The expected time delay, under conditions of real traffic.

[16] Tanenbaum A.S., Computer Networks, Prentice-Hall, third edition, 1996.

[17] Wolisz A., Fitzek F.H.P., QOS support in wireless networks using simultaneous MAC packet transmission (SMPT), ATS, April 1999.

[18] IRCACHE, The NLANR Web Caching project, ftp.ircache.net, www.ircache.net/FAQ/.