Analysis of transient of the datagram traffic for a demand assignment TDMA satellite access scheme

Nedo Celandroni

CNUCE, Institute of National Research Council
Via S. Maria 36 - 56126 Pisa - Italy
Phone: +39-50-593207/593312/593203
Fax: +39-50-904052- Telex: 500371
E-mail: n.celandroni@cnuce.cnr.it

Abstract - A modified version of the datagram capacity assignment of the FODA access scheme, named FODA/IBEA, is briefly presented. The main difference from the previous version (besides the fade countermeasure feature) is the possibility of the system to migrate, depending on the overall loading conditions, from a sort of fixed TDMA (called pre-assignment mode), in which each active station shares the spare capacity with all the others, to a more complex mechanism. The response of the system to a step of traffic at one of the active stations is studied, and the analytical expressions of the most significant variables are derived. The results of the analysis are first compared with the experimental ones and then used to show the sensitivity of the system's behavior with respect to some parameters. The possibility of adopting the capacity allocation algorithm in a distributed mode is also discussed.

Key words: Demand assignment TDMA, Satellite access scheme, FODA/IBEA, Anisochronous traffic, Transient analysis

1. Introduction

In [1] the FODA (Fifo Ordered Demand Assignment) satellite access scheme in TDMA (Time Division Multiple Access) was presented, together with the performance obtained with the simulation study and the experimental results of three prototype units on satellite, working at a constant speed of 2 Mb/s. This system was designed to optimally share the capacity of a geosynchronous satellite bent pipe channel among many stations on a demand-assignment basis. Both isochronous (*stream*) and anisochronous (*datagram*) traffic were supported. FODA was then modified to support rain fade countermeasure techniques based on the adaptation of the energy per information bit to each individual link status, which depends on atmospheric conditions. This made the system particularly suitable for employment in Ka band, which is heavily affected by rain fading. In the new system the total attenuation of each link (up-link plus down-link) is compensated for by varying the transmission power, coding and bit rates. A multi-channel TDMA access is envisaged to fully exploit the satellite transponder capacity. The transmission power variation must therefore ensure a constant back-off at the transponder input, to avoid excessive intermodulation noise. The power control can thus be used to compensate up-link attenuations only, while the total compensation is completed by varying the coding and bit rates as well. The name of the new system is FODA/IBEA (Information Bit Energy Adaptive). The presentation of the system and the performance of the fade countermeasure mechanism can be found in [2, 5]. The performance of the prototype, in terms of jitter on stream data and datagram assignment algorithm behavior, is outlined in [3] in which the results of measurements obtained on the ITALSAT satellite with four working stations are shown. The datagram assignment algorithm behavior was investigated with simulations, using four different traffic models: fixed rate, Poisson, two-state Markov modulated Poisson and fractal. The simulation results reported in [4] show that the most critical traffic type is Markov modulated Poisson with time parameters that force the system to work permanently in a transient condition. Demand assignment schemes for the capacity allocation of geosynchronous satellite channels suffer from a considerable delay between the request and the allocation. To combat the effect of this delay, a model of the system was derived which was very helpful when designing the system. The aim of the present paper is the presentation of this model to study the effect of a step of traffic at one of the stations. In order to select the above mentioned event alone, all the other stations are assumed to be in steady state conditions and loaded with fixed rate traffic. Under linearity conditions we will see that the step of traffic only affects the station which is loaded with that step. The transient effects, in terms of delay and queue length, as well as the sensitivity relevant to certain system parameters can thus be analyzed simply by considering the station that experiences the transient.

For the sake of comparison with FODA/IBEA, some other access schemes are referenced [7-13].

In Section 2 of this paper, the datagram assignment algorithm is briefly described, including a recently adopted feature called *pre-assignment mode,* which significantly improves the end-to-end delay during the transient, when one of the active stations passes from an unloaded condition to a constant rate traffic. In Section 3 the system is modeled and the analytical expressions, which give the temporal functions of the queue length and of the end-to-end delay, are derived using the Laplace transform method. In Section 4 the analytical results are compared with the values measured using the system prototype on the satellite and some results from the analysis are shown to justify some system design features. The adoption of a distributed algorithm for the capacity assignment is considered and the performance is compared with the centralized algorithm as well. Conclusions are drawn in Section 5.

2. The datagram assignment algorithm

A master station is responsible for system synchronization and for the capacity allocation on request from the traffic stations. These tasks are accomplished by sending a reference burst which contains the burst time plan (BTP), at the beginning of each 20 ms *frame*. The BTP is the transmission window layout allocated on the traffic stations' request basis. Only one window is allocated for each requesting station, in order to save the overhead due to the rather long preambles needed by the modem for each burst synchronization. Inside the transmission window the traffic stations multiplex data coming from different applications of both stream and datagram type. Data packets generated by each application are sent in sub-bursts, adapting the coding and bit rates to the current condition of each individual link and to the BER required by each application. Broadcast and multicast transmissions are also supported. In this case the worst link status is considered to adapt the transmission parameters. The transmission bit rate can vary from 1 to 8 Mb/s and the coding rates used are: 4/5, 2/3, 1/2 and uncoded.

While the delay of the stream traffic must be kept as low as possible and with constant packet inter-arrival times, the delay of the datagram traffic is not so critical. In any case it must be investigated in order either to tune-up efficient application protocols or to implement suitable congestion control mechanisms, which are required to improve system stability.

Stream capacity is guaranteed by the system, i.e. once the request has been accepted the assignment is maintained until released by the user. The total stream capacity is allocated up to a maximum value; the remaining amount of the capacity is devoted to the datagram. The algorithm for the datagram capacity allocation does not depend on the presence of stream data in the frame.

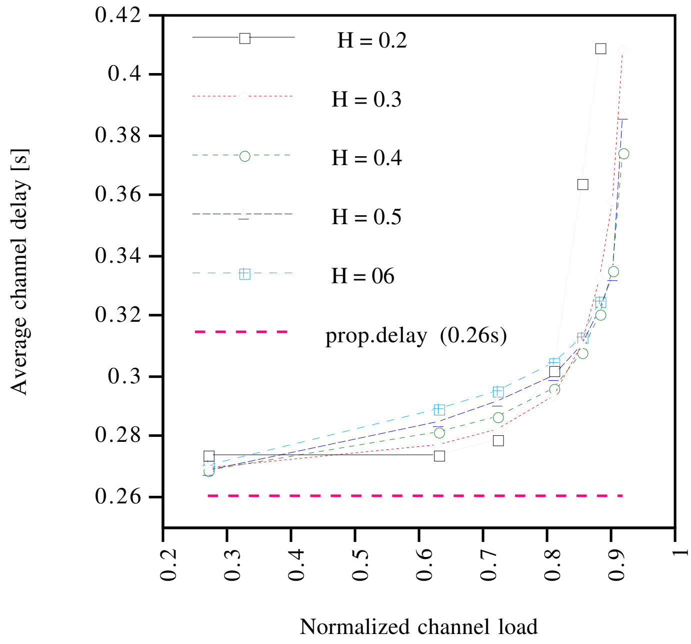


Fig. 1. Average end-to-end delay (10 stations), versus channel load, for various values of H. The system is loaded with Poisson generators. The channel load includes the packet headers.

The request for a datagram transmission window, sent by a traffic station, is proportional to the traffic i coming into the station plus the backlog q, i.e. the volume of data waiting for transmission on satellite. We have:

$$r = q + H\, i \qquad (1)$$

where H is a temporal constant of proportionality. Simulation results, obtained by loading the channel with Poisson generators of datagram traffic for 10 stations (Fig. 1), suggested using 0.4 s as the best compromise for the H parameter. This value, in fact, gives the best performance at high channel loads, while it does not significantly increase the delay, at low-medium loads, with respect to lower values. Datagram requests are issued as frequently as possible, to update the master station about the latest changes in the traffic situation. The requests are piggy-backed with the data, when an assignment already exists for the requesting station, otherwise they are sent by using a control window whose assignment is guaranteed with a minimum frequency.

The master organizes the requests of all the traffic stations into a *ring,* which it scans cyclically to compute the assignments. Any datagram request received from the same station replaces the previous value. The length of the assigned transmission window w is proportional to the request in a range of values between a minimum $(w_{\min})$ and a maximum $(w_{\max})$. We have:

$$w = \min(w_{\max},\ \max(w_{\min},\ f\ \ r)). \qquad (2)$$

The coefficient f has been chosen, for the current implementation, equal to the number of active stations N divided by 100, with a minimum of 0.05 and a maximum of 0.5. The threshold $w_{\min}$ has been introduced for efficiency purposes. It prevents the information part of the allocation from being too small in comparison with the transmission overhead, due to preambles and headers. The threshold $w_{\max}$ prevents a heavy loaded station from removing too much capacity from the other stations.

After each assignment, the datagram request is decreased by the assignment itself, and the next request is analyzed, if space is still available in the frame. The first assignment that does not entirely fit the current frame is considered as first in the next frame, where the rest of the computed amount is assigned. All the space up to the end of the frame (if insufficient for a *minimum assignment*) is given as an over-assignment to the last processed station.

The ring is not scanned more than once in a frame. We call *assignment cycle* a complete scan of the ring. Thus, no more than one assignment cycle is made in a frame. If space is still available in the frame after an entire assignment cycle, that space is shared among all the active stations, including those stations which had no datagram assignment in that frame.

The datagram assignment algorithm was studied to allow a quick response time to traffic spikes at the station input. When the system is lightly loaded, its behavior approaches that of a fixed TDMA. This situation is called *pre-assignment* mode. The system gradually migrates towards a "pure" FODA assignment scheme when the channel load increases beyond a certain limit, called *pre-assignment limit.* A moderately loaded system can absorb abrupt traffic variations without appreciable delays, because each traffic station has some spare capacity.

In pre-assignment mode the assignment cycle is always one frame long, and some spare capacity is always available to be shared among all the stations. When the assignment cycle exceeds one frame, capacity is no longer available to be shared.

3. Linear analysis

For the linear analysis of the system let us make the following assumptions. Only the station considered has a variation of the input traffic at the time $t=0$, which is assumed as the reference time. The station may have a pre-existent constant rate traffic. All the other stations run with a constant rate traffic and the whole system is in a steady state condition at the time $t=0$. All the variables considered are assumed to be continuous as opposed to discrete, like they are in reality. In particular, the allocations are assumed to be uniformly distributed in time, so no effect due to the duty cycle less than one is considered and the concept of frame is not taken into account. The allocation quantities must lie in the linear zone between the lower and the upper thresholds of the assignment.

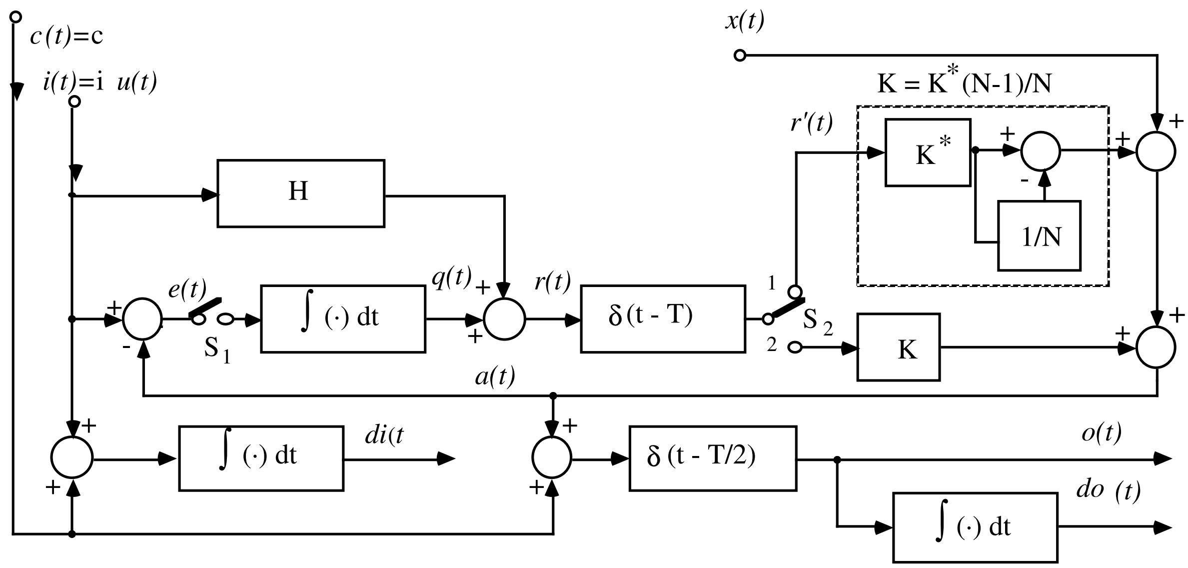


Fig. 2 The system block diagram.

The system model that is valid in linearity conditions is depicted in Fig. 2, where:

- $u(t)$ is the unitary step function: $u(t) = 0$, for $t < 0$; $u(t) = 1$, for $t \geq 0$;
- $i(t) = \mathrm{i}\ u(t)$ is the input traffic transient function, i is the amplitude of the step;

- $c(t)$= c is the pre-existent fixed rate traffic;
- $\delta(t-\alpha)$ is a delay operator giving the output $f(t-\alpha)$ when solicited with an input function $f(t)$. It models the effect of the satellite channel delay;
- T is the interval of time between the request sending time and the allocation receiving time. It has been approximated for simplicity to a constant value equal to the duration time of a double round-trip plus a frame. The transfer time, i.e. the satellite round-trip-time plus some overheads, has been assumed to be $T/2$;
- $q(t)$ is the queue of the data waiting to be transmitted, i.e. the backlog;
- $a(t)$ is the part of the transmitted traffic that relates to the transient, while $q(t) \geq 0$;
- $e(t)$ = $i(t)$ - $a(t)$ is the traffic not transmitted. It is responsible for the queue variation;
- $o(t)$ is the total traffic that has been transmitted to and has arrived at the destination station, i.e. $c(t)+a(t)$ delayed by a round trip;
- $d_i(t)$ is the total amount of data that has entered the system since the time t = 0;
- $d_o(t)$ is the total amount of data that has arrived at the destination station since the time t = 0;
- S_1 is considered closed while $q(t)$ > 0, and open when $q(t)$ becomes 0. It is considered closed again when $e(t)$ becomes positive. While S_1 is open the assignment is greater than the request;
- S_2 selects the system working in pre-assignment mode (position 1) or non pre-assignment mode (position 2);
- H is the factor that determines the weight of the traffic in the user request;
- $r(t) = q(t) + H \ i(t)$ is the user request and $r'(t)$ is the function $r(t)$ delayed by the time T to take into account the delay between the request and the assignment;
- K is the assignment coefficient. The master station assigns a fraction f of the user request whenever the requesting station is scheduled. We thus have:

  $$K = f / T_a, \qquad (3)$$

  where T_a is the allocation cycle duration, i. e. the time between two consecutive allocations to the same station. When the system works in pre-assignment mode the assignment is made in each frame, so $T_a = T_f$. We denote by

  $$K^* = f / T_f \qquad (4)$$

  the assignment coefficient in pre-assignment mode;

- $x(t)$ is the resulting over-assigned capacity while S_1 is closed. It is the sum of the pre-assigned capacity due to the redistribution of the spare capacity, denoted by $p(t)$, and the contribution due to the constant traffic before the transient, which is the difference between the previous assignment and the capacity used by the pre-existing constant traffic. We have

  $x(t) = p(t) + (HK^* - 1)\ \mathrm{c}$ in pre-assignment mode ($p(t) > 0$), or (5)

  $x(t) = (HK - 1)\ \mathrm{c}$ in non pre-assignment mode ($p(t) = 0$)

  The capacity $x(t)$ is assigned since the time t=0. The contribution due to the pre-existing traffic was thus considered separately from the transient step. It is assumed that $x(t) \geq 0$;

In pre-assignment mode, the residual capacity $C_r(t)$ is divided among all the N active stations, thus:

$$p(t) = C_r(t) / N,$$

When the station considered receives the assignment for the transient traffic $s(t) = K^* r'(t)$, $C_r(t)$ becomes $C_r(t) = \mathrm{C_r} - a(t)$, where $\mathrm{C_r}$ is the residual capacity before the transient.

We thus have

$$a(t) = (\mathrm{C_r} - K^* r'(t)) / N + (HK^* - 1)\mathrm{c} + K^* r'(t), \quad \text{from which}$$

$$a(t) = \mathrm{p} + (HK \frac{N}{N-1} - 1)\mathrm{c} + K\ \ r'(t), \text{ where:}$$

$$K = K^*(1 - 1/N), \tag{6}$$

$$\mathrm{p} = \mathrm{C_r}/\mathrm{N}.$$

The model is thus simplified because it is the same as considering the pre-assigned capacity equal to a constant p and an assignment coefficient K given by (6), as depicted in Fig. 2.

The linearity conditions are verified provided that the assignment for the transient traffic is

$$s(t) \leq \mathrm{C_r} + (HK^* - 1)\mathrm{c}. \tag{7}$$

for all the transient duration. In this case all the other stations are not affected by the transient, because it is absorbed by the spare capacity and the station considered receives an assignment proportional to the request.

In non-pre-assignment mode $\mathrm{C_r} = 0$ and T_a has the form

$$T_a = [f \sum_1^N (q_n + H\ \ i_n)] / \mathrm{C_d}, \tag{8}$$

where $\mathrm{C_d}$ is the system capacity reserved for datagram, q_n and i_n are the queue and the traffic at the n^{th} station, respectively. Substituting (8) in (3) we have:

$$K = \mathrm{C_d} / \sum_{1}^{N} (q_n + H \ i_n). \tag{9}$$

In this case the linearity conditions are not rigorously satisfied and K is generally a variable coefficient. However, if the amplitude of the transient traffic is sufficiently small with respect to the sum of all the other requests, K can be approximated to a constant and the analysis still gives valid results.

Looking at Fig. 2 we can write the following system, while S_1 is closed

$$\begin{aligned}
&x(t) = \mathrm{p} + (HK^* - 1)\ \mathrm{c}, && \text{or:}\\
&x(t) = (HK - 1)\ \mathrm{c}, && \text{for} \quad \mathrm{p} = 0\\
&a(t) = u(t-T)\ [K\int_0^{t-T} [i(t-T) - a(t-T)]\ d(t-T) + HKi(t-T) + x(t) &&\\
&q(t) = \int_0^t [i(t) - a(t)]dt &&\\
&d_i(t) = \int_0^t [c(t) + i(t)]dt &&\\
&o(t) = u(t-T/2)[c(t-T/2) + a(t-T/2)] &&\\
&d_o(t) = \int_0^t o(t)dt. &&
\end{aligned} \tag{10}$$

Denoting with capital letters the Laplace transforms of the temporal functions and assuming zero initial conditions, the transforms of the system (10) become

$$\begin{aligned}
&X(s) = \mathrm{p}/s + (HK^* - 1)\ \mathrm{c}/s, && \text{or :}\\
&X(s) = (HK - 1)\ \mathrm{c}/s, && \text{for p=0}\\
&A(s) = \frac{[1 + sH]Ke^{-sT} I(s) + sX(s)}{s + Ke^{-sT}} &&\\
&Q(s) = \frac{[1 - HKe^{-sT}]I(s) - X(s)}{s + Ke^{-sT}} &&\\
&D_i(s) = \frac{1}{s}[C(s) + I(s)] &&\\
&O(s) = e^{-sT/2}C(s) + \frac{[1 + sH]Ke^{-s(T+T/2)} I(s) + e^{-sT/2} sX(s)}{s + Ke^{-sT}} &&\\
&D_o(s) = \frac{O(s)}{s}. &&
\end{aligned} \tag{11}$$

Considering the case of a step of input traffic with amplitude i, beginning at the instant $t=0$, we have $I(s)=\mathrm{i}/s$.

Applying the final value theorem:

$$\lim_{t\to\infty} f(t)=\lim_{s\to 0} s\mathbf{L}\{\mathrm{f(t)}\}$$

to the third of (11), we get the steady state solution for the queue length

$$\lim_{t\to\infty} q(t)=\frac{(1-HK)\mathrm{i}-\mathrm{p}-(\mathrm{HK}^*-1)\mathrm{c}}{K} \qquad \text{and} \tag{12}$$

$$\lim_{t\to\infty} q(t)=\frac{(1-HK)(\mathrm{i}+\mathrm{c})}{K} \qquad \text{for p = 0.}$$

In non-pre-assignment mode (p = 0) it is possible to get a null queue and a null queuing delay, when steady state is reached, if

$$HK\geq 1, \tag{13}$$

so, substituting K given by (9) in (13), we get

$$H\ \mathrm{C_d}/\sum_1^N (q_n+Hi_n)\geq 1.$$

Considering null q_n for each n, the above condition becomes

$$\sum_1^N i_n \leq \mathrm{C_d}, \tag{14}$$

as already seen in [1].

For the pre-assignment mode (p > 0), let us replace the notation of i and c (the transient and the constant traffic parts of the considered station) with i_{1t} and $\mathrm{i}_{1\mathrm{c}}$, respectively. So we have: $i_1=\mathrm{i}+\mathrm{c}=i_{1t}+i_{1c}$. The pre-assigned capacity p can thus be expressed as

$$\mathrm{p}=\{\mathrm{C_d}-K^*[Hi_{1c}+q_1+\sum_2^N (q_n+Hi_n)]\}/N\cdot \tag{15}$$

Substituting p, given by (15), in the first relation of (12), the condition to get a null queue becomes

$$(1-HK)i_{1t}\leq \{\mathrm{C_d}-K^*[Hi_{1c}+\sum_2^N (q_n+Hi_n)]\}/N-(1-HK^*)i_{1c} \tag{16}$$

Substituting (7) and (4) in (16) and considering q_n null for each n, (16) becomes

$$(1-\frac{fH}{T_f})i_1 \le (C_d - \frac{fH}{T_f}\sum_1^N i_n)/N,$$

from which we see that, if

$$\frac{fH}{T_f} = 1, \qquad (17)$$

i_1 has no limitation, provided that the condition (14) is respected.

If $fH/T_f > 1$ the system works in pre-assignment mode while

$$\sum_1^N i_n \le \frac{T_f}{fH}\mathrm{C_d} \qquad (18).$$

When the load increases beyond the pre-assignment limit given by (18), the system works in non-pre-assignment mode and the product HK^* tends to unity as the system load tends to maximum, i.e. when the condition (14) tends to equality.

The condition $fH/T_f < 1$ is peculiar and more restrictive for the linear analysis validity. The system always works in pre-assignment mode, but as the load tends to the maximum, i.e. when the condition (14) tends to equality, it must be: $i_1 \le \mathrm{C_d}/N$, so: $i_{1t} \le \mathrm{C_d}/N - i_{1c}$.

It seems that the condition $fH/T_f = 1$ is the best, because each station has a pre-assigned $1/N$ of spare capacity to absorb moderate transients of traffic immediately, while, in case of greater transients, up to the entire spare capacity is devoted to the requesting station by the allocation mechanism. However, as it is reasonable to think that each station traffic has an order of magnitude proportional to $1/N$, when the system works with a high number of stations the condition $fH/T_f = 1$ may cause too many allocations to fall below the lower threshold of the assignments with a consequent compression of system dynamics and a waste of capacity. That is why we chose to make fH proportional to N. More precisely, we chose to fix the factor H as the best result from the simulation and $f = N/100$ with a $f_{\min} = 0.05$ and $f_{\max} = 0.5$. With $H = 0.4$ and $T_f = 0.02$ the result is always $fH/T_f \ge 1$.

The above analysis is valid for $q(t) \ge 0$. After that the model changes and S_1 opens. However, in order to find the conditions for $\underset{t\to\infty}{q(t)} = 0$, the analysis is still valid, because once $q(t)$ reaches the zero value, under the conditions (7) or (13), it remains null.

The temporal functions which express system behavior during the transient are obtained by inverting the transforms (11) (see Appendix). We have

$$q(t)=\sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{(\mathrm{i}-\mathrm{x})[t-(j-1)T]^{j}u(t-(j-1)T)K^{j-1}-\mathrm{i}HK^{j}(t-jT)^{j}u(t-jT)\}$$

$$d_i(t)=(\mathrm{c}+\mathrm{i})\ t$$

$$o(t)=u(t-\frac{T}{2})\mathrm{c}+\sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{\mathrm{i}K^{j}(t-jT-\frac{T}{2})^{j-1}[t+(H-T)j-\frac{T}{2}]u(t-jT-\frac{T}{2})+ \qquad (19)$$

$$\mathrm{x}\ K^{j-1}j[t-(j-1)T-\frac{T}{2}]^{j-1}u(t-(j-1)T-\frac{T}{2})\}$$

$$d_o(t)=u(t-\frac{T}{2})(t-\frac{T}{2})+\sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{\mathrm{i}K^{j}[\frac{(t-jT-\frac{T}{2})^{j+1}}{j+1}+H((t-jT-\frac{T}{2})^{j}]u(t-jT-\frac{T}{2})]+$$

$$\mathrm{x}\ K^{j-1}(t-(j-1)T-\frac{T}{2})^{j}u(t-(j-1)T-\frac{T}{2})\}$$

where:

$\mathrm{x}=\mathrm{p}+(HK^{*}-1)\ \mathrm{c}$ or

$\mathrm{x}=(HK-1)\mathrm{c}$ for p = 0.

Expressions (19) contain summations of infinite terms, but these summations can be truncated at the j^{th} term without losing any precision within the interval of time: $0 \le t \le jT$. This method is particularly suitable for analyzing temporal functions over limited intervals of time such as transients.

The end-to-end delay experienced by the data in the satellite network crossing is the sum of the queuing delay and transfer delay. It can be computed as the difference $\tau - t$, where τ is such that:

$$d_o(\tau)=d_i(t). \qquad (20)$$

To plot the delay as a function of t may be difficult. Although getting $d_i(t)$ is straightforward, from the second relation of (19), the solution of equation (20) is complex for high values of j. It is thus preferable to plot the inverse function of the delay, i.e., after fixing a value of τ, by computing $d_o(\tau)$ directly, from the last relation of (19) and then computing t as the inverse function of $d_i(t)$. Substituting the second relation of (19) in (20), we have

$$t=\frac{d_o(\tau)}{c+i}.$$

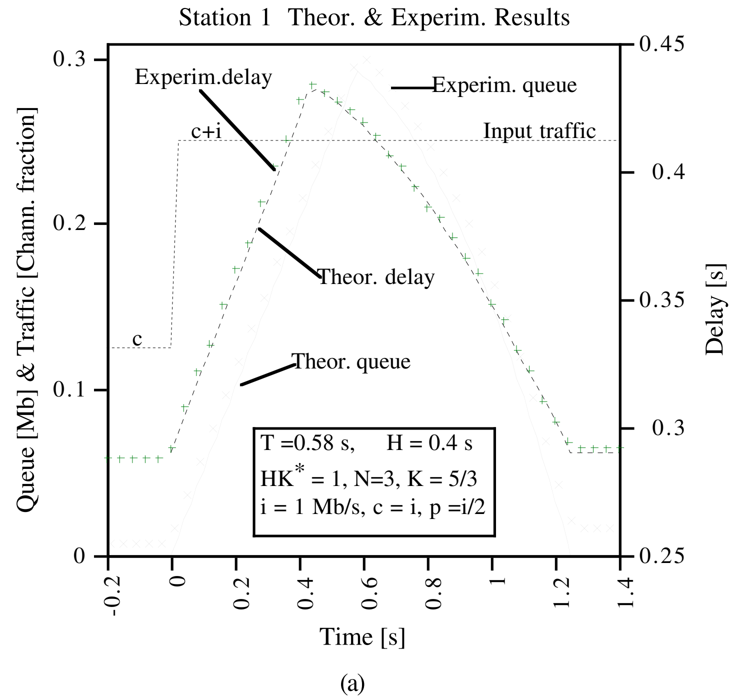
4. Results

4.1. Model validation

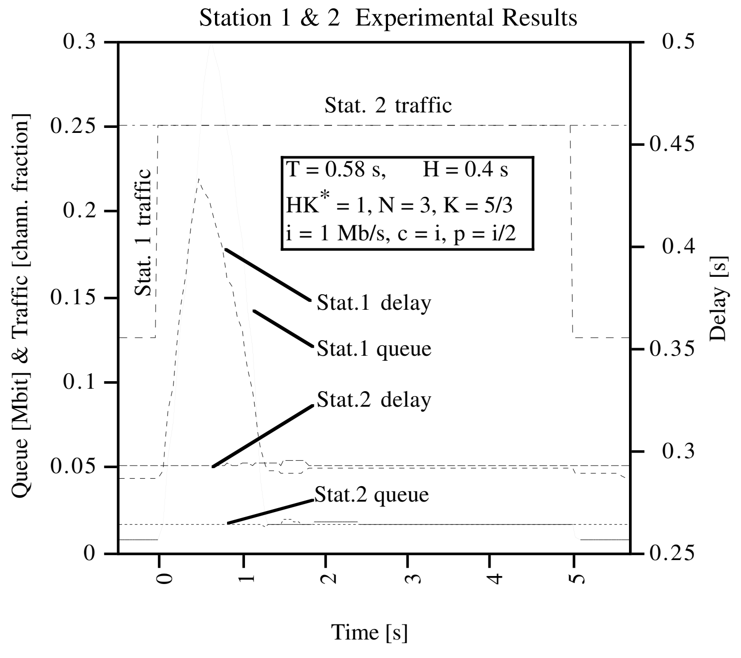
To check the validity of the present analysis a comparison was made between the analytical and experimental results. In Figs. 3-5 the queue length and the delay as a function of the time are reported, together with the traffic of station 1 which experiences the transient for three different loading conditions of the system. For two loading conditions (Figs. 3 and 4) a comparison between stations 1 and 2 of the measured queue and delay are reported as well. Three active stations were used for the tests. The values measured are averaged over the interval of time indicated by the symbols, to eliminate the superposition of the saw-tooth due to data traffic packetization, the transmitting duty cycle, and frame effects. This allows an easier comparison between the analytical and the experimental results.

In Fig. 3 condition (7) for the linear analysis validity is fully respected. The results are practically in total agreement. The measured values are slightly higher than the theoretical ones, because the duty cycle and the frame effects are not considered in the analysis. In this case the assignment to station 1 is fully proportional to the request for all the transient duration, according to our model. The other stations are not affected at all by station 1's transient as can be observed in Fig. 3b. In the case shown in Fig. 4, condition (7) is not respected all the time. In fact, a small hump can be observed, on the values measured at the station 1, 2 s from the beginning. This is due to the system reacting to the transient. For a while the assignment cycle T_a becomes bigger than T_f, causing a drop in the K coefficient. This effect is not considered in our analysis. The other stations are moderately affected by station 1's transient (Fig. 4b). In the case shown in Fig. 5 we are well beyond the validity of the linear analysis. In this case the queue and the delay eventually diverge. We attempted to make the analytical results fit, using a K coefficient of 1.4. Initially a certain agreement is observed, but after the system has reacted the results diverge as expected.

In Fig. 6 all the quantities expressed by the relations (19) are plotted, for the same conditions as Fig. 3.

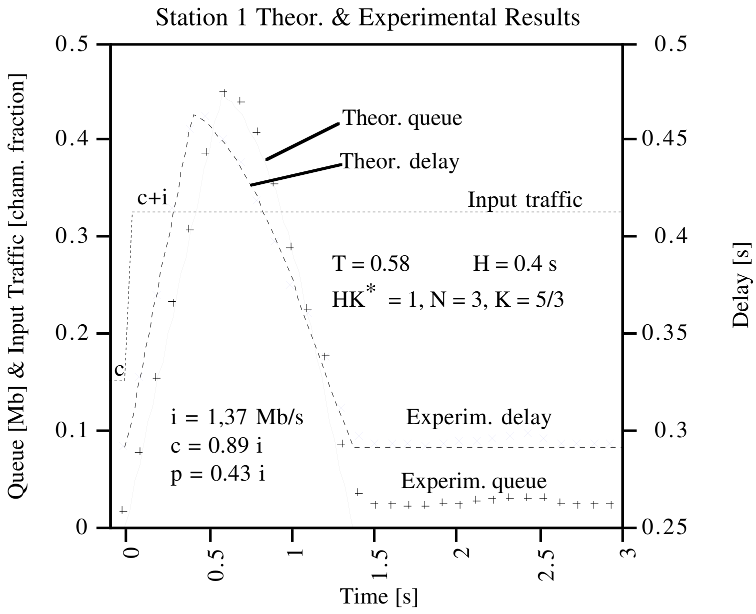


(a)

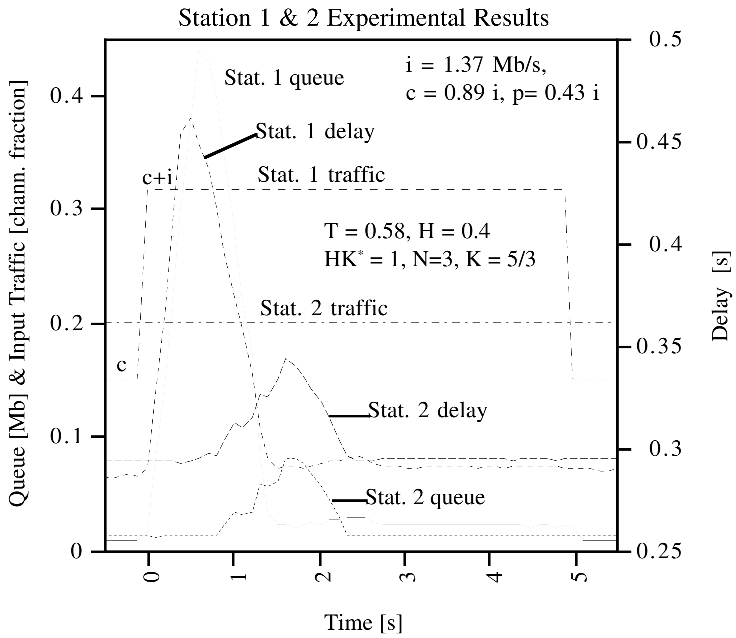


(b)

Fig. 3. Queue and delay versus time during a transient due to a 1 Mb/s traffic step at one of the three working stations. a) Comparison between the analytical and the experimental results at station 1. b) Comparison between experimental values at station 1 and 2. The linearity condition is fully respected.



(a)



(b)

Fig 4. Queue and delay versus time during a transient due to a 1.37 Mb/s traffic step at one of the three working stations. a) Comparison between the analytical and the experimental results at station 1. b) Comparison between experimental values at stations 1 and 2. The linearity condition is not respected for the entire transient.

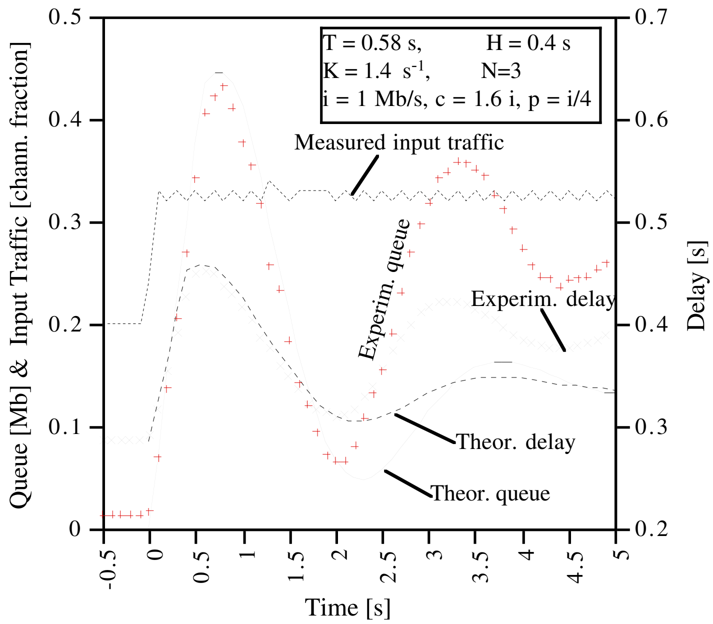


Fig. 5. Queue and delay versus time during a transient due to a 1 Mb/s traffic step at one of the three working stations. The linearity condition is not respected.

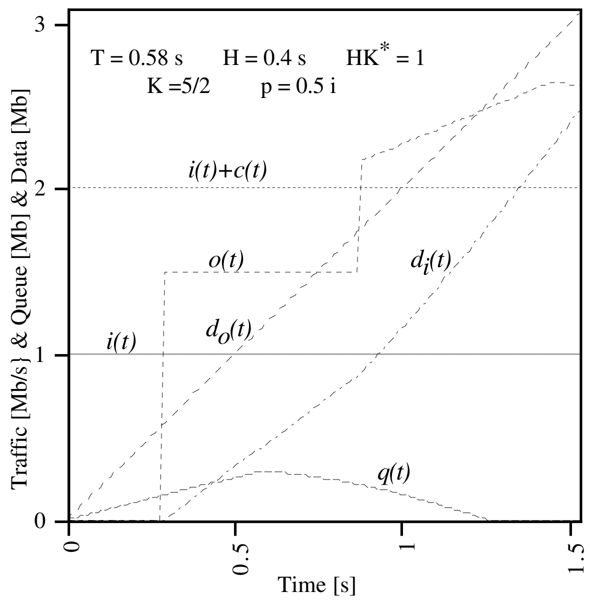
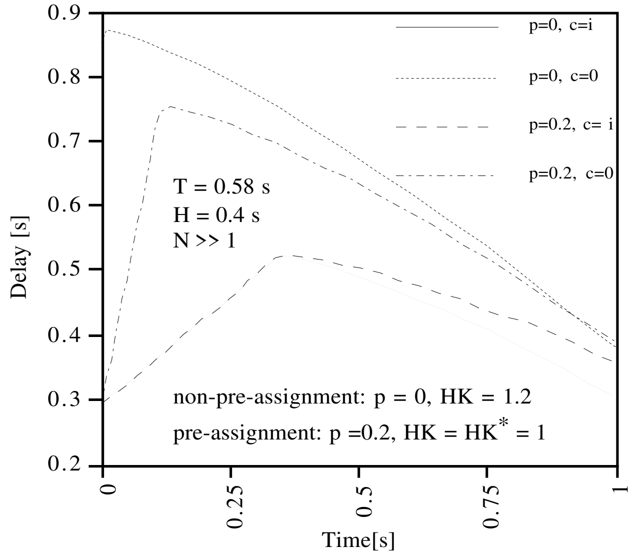


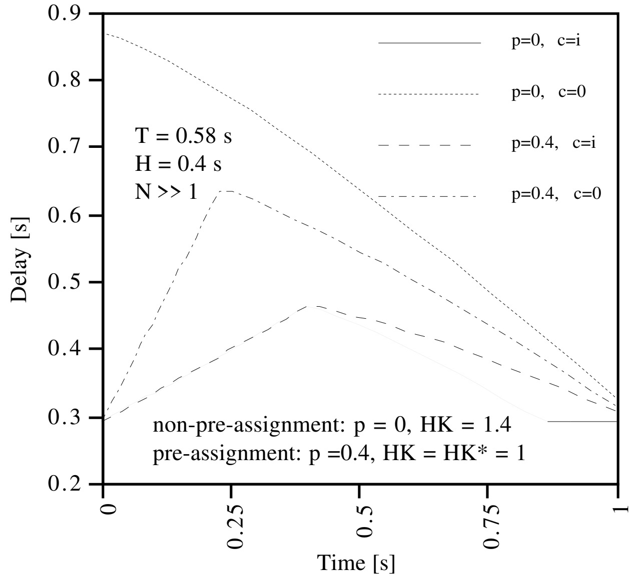
Fig. 6. All the analytically derived temporal functions, during the transient case in Fig 3.

4.2. Model applications

Considering the field in which the linear analysis is valid, let us now look at some comparative results which are useful for sizing the system parameters.



(a)



(b)

Fig. 7. Comparison between pre-assignment and non-pre-assignment modes, for both pre-loaded and non pre-loaded stations. The delay is plotted as a function of the time. *a*) the system is 83% full. *b*) the system is 71% full.

In Fig. 7 the delay-versus-time curves are compared with the system working in pre and non-pre-assignment modes, respectively. These curves highlight the importance of working in pre-assignment mode. In fact, the gain in the non-pre-assignment case when the station works with a pre-existing traffic, equal in size to the transient step (c = i), is very small, while the maximum delay in the pre-assignment case, when the station is not loaded at the transient step time (c = 0) is significantly lower than the one experienced in the pre-assignment case. This result becomes more evident with a lower load in the system (case *b*). Note that the maximum delay and the ratio between the maximum and the average delay are parameters which significantly affect the performance of applications that use datagram such as bulk data transfer.

In Figs. 8-10 the delay-versus-time curves are plotted for different values of the H factor in non-pre-assignment mode. Figures 8 and 9 refer to the system when it is loaded to the maximum ($H\ K$=1), and the station is working with c = i and c = 0, respectively. On the other hand, Fig. 10 shows a less loaded system ($H\ K$ = 1.3); the delay curves are plotted for two values of the H factor. The delay improves with lower values of H, which means higher values of K, as the product $H\ K$ is constant for a certain loading condition. However, the linear analysis does not consider system efficiency. A higher value of K, from (3), means a shorter assignment cycle T_a. As the overhead due to the sum of the preambles is the same for the same number of stations, the efficiency worsens with lower values of H. This behavior is well confirmed by system simulation. Figure 1 shows that, for a low load in the system, the delay improves with H decreasing, in agreement with the linear analysis, while it improves with H increasing when the system approaches saturation and the efficiency influence becomes predominant.

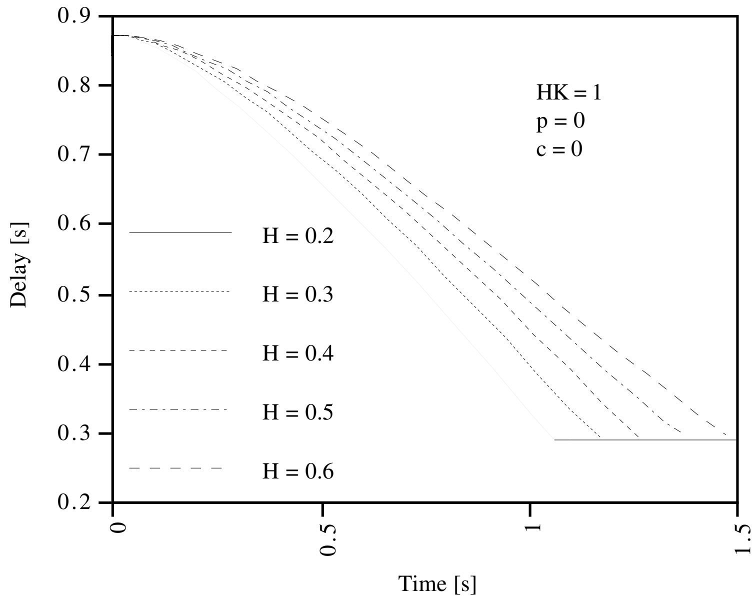


Fig. 8. Comparison of the delay-versus-time curves, during the transient, for different values of H, in non-pre-assignment mode. The station is not loaded before the traffic step. The system is fully loaded.

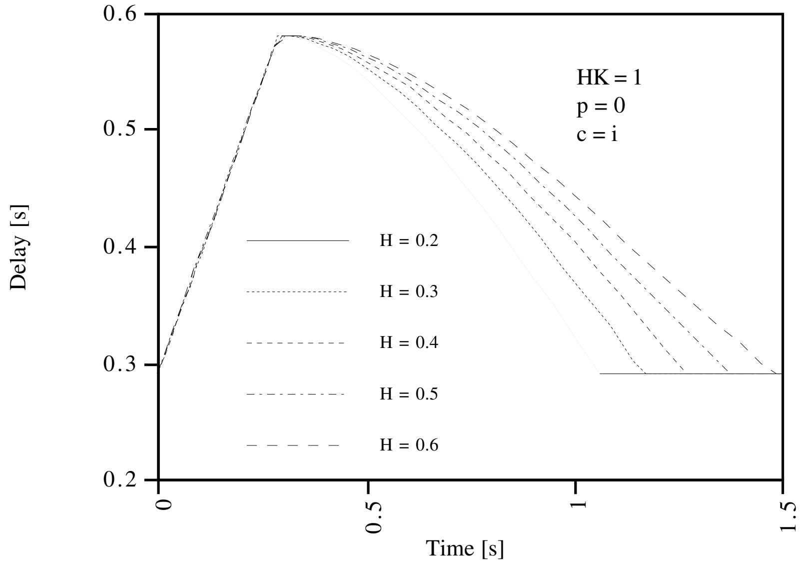


Fig. 9. Comparison of the delay-versus-time curves, during the transient, for different values of H, in non-pre-assignment mode. At the traffic step time the station has a load of the same amplitude as the traffic step. The system is fully loaded.

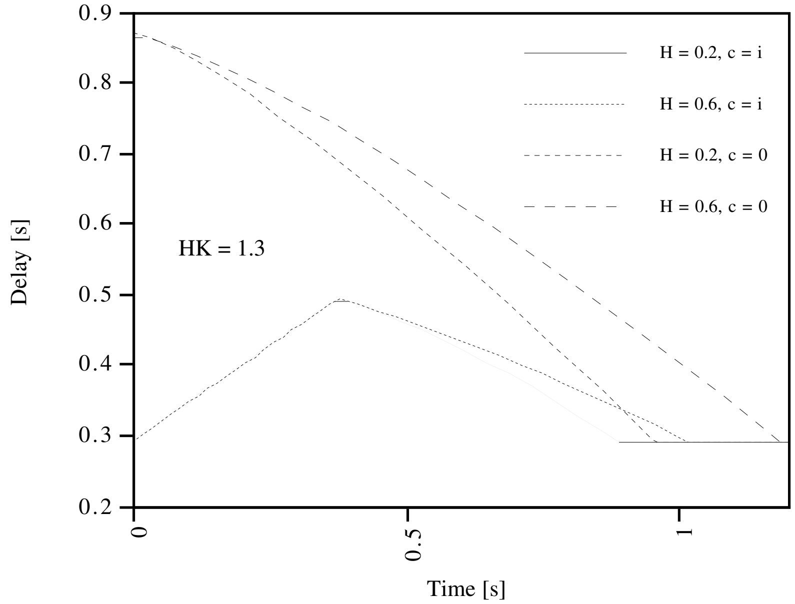


Fig. 10. Comparison of the delay-versus-time curves, during the transient, for two values of H, in non-pre-assignment mode, for both non-pre-loaded and pre-loaded station cases. The system is 72% full.

So far we have considered the system working in centralized mode. This means that the master station is responsible for both the system synchronization and the channel capacity assignment. This technique gives good stability to the system, but introduces a delay between request sending and assignment reception in the order of two round trips (T parameter). The introduction of a distributed technique for the capacity assignment is very attractive, because in this case the parameter T is practically halved. In the solution proposed in [6] the master station is still responsible for system synchronization, while each station computes its own assignment of the capacity by considering the requests of all the stations, which are sent in the signaling messages. In this way one round trip time is saved in the request-allocation delay. Figure 11 shows the considerable gain during the transient in the performance of the distributed case with respect to the centralized one.

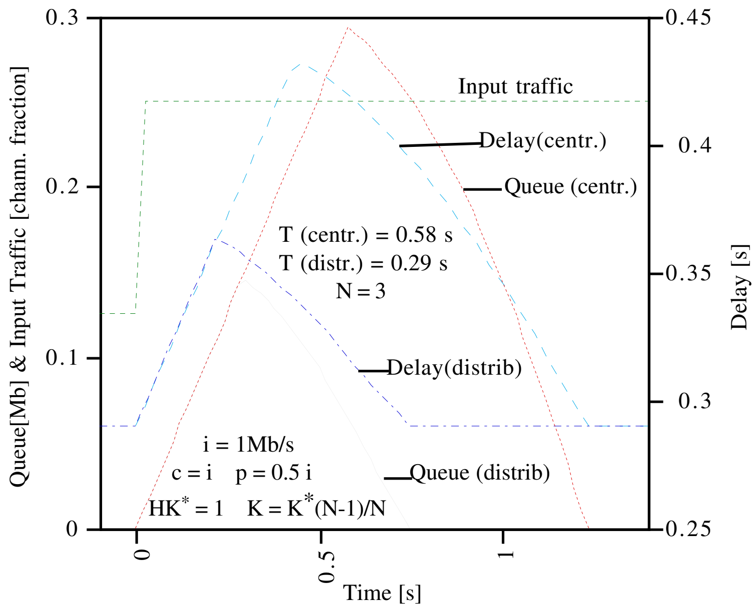


Fig. 11. Comparison between centralized and distributed modes of the capacity allocation. The queue and the delay are plotted as functions of time during the transient, due to a step of traffic at one of the three working stations. At the traffic step time the station has a load of the same amplitude as the traffic step.

The drawback of the distributed systems is the possible worsening of the overall stability due to the higher probability of losing signaling messages. In fact, each station is able to compute its assignment only if the reception of all the other stations' requests is correct. If there are losses the station is not allowed to transmit for one frame. This problem may become significant if the transmitting signal fades, due to bad atmospheric conditions, which reduces the Eb/No (bit energy over one-sided noise power spectral density) available at each station's receiver. This problem is studied in [6] and a solution is given, based on making the signaling messages redundant. The probability of the loss of signaling messages is evaluated for various numbers of active stations, considering the link attenuation distribution probability. Techniques to recover from situations caused by signaling message loss are studied as well.

Minimum Eb/No [dB]	$n_a = 2$	$n_a = 3$	$n_a = 4$
7	$4\ 10^{-5}$	$4\ 10^{-7}$	$4\ 10^{-9}$
8	$1.7\ 10^{-6}$	$3\ 10^{-9}$	$5\ 10^{-12}$

Tab. 1. Signaling message loss probability for different Eb/No ratios and redundancy coefficients.

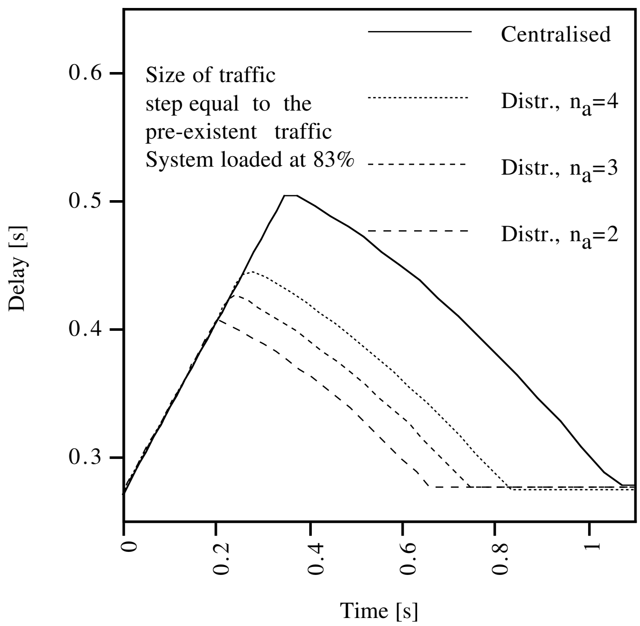


Fig. 12. Delay versus time curves, during a traffic step, for both centralized and distributed systems, with different values of n_a. The system is loaded at 83%, N>>1, and the station which experiences the transient is pre-loaded with a constant rate traffic.

The most significant results given in [6] are summarized in Tab. 1, in which the probability that one of the 32 stations considered loses one of the allocation requests is given for two values of the minimum operating Eb/No and three values on the signaling redundancy factor n_a. The coefficient n_a indicates the number of consecutive frames in which every station sends the same allocation request. The effect of this procedure is to freeze the allocation computing mechanism for n_a frames, thus increasing the parameter T. In Fig. 12 the delay versus time during the transient, for the distributed system, is reported as a function of the parameter n_a, and compared with the centralized case. We can see that considering the factor $n_a = 3$, which gives a very acceptable loss probability even at an Eb/No of 7 dB, the performance gain of the distributed system with respect to the centralized one is still considerable. This result is confirmed by the complete simulation study for the performance comparison between the datagram assignment algorithms of the two types of system reported in [4].

5. Conclusions

The datagram capacity assignment of the FODA/IBEA access scheme has been described. An analytical study of the system working in linearity conditions has been carried out to investigate transient behavior, when a step of traffic is applied to one station. The analytical results have been compared with the experimental ones, showing good agreement under certain conditions.

The analytical results have been used to show the improvement in the end-to-end delay working in pre-assignment mode, when the station considered has no load before the traffic step. Problems of system efficiency lead, however, to a compromise in sizing the parameters that determine the limit of the pre-assignment mode (f factor) and the weight of the traffic in the user request (H factor). To tune up these factors the results of system simulation have been taken into account.

The possibility that the system may work with the capacity allocation algorithm distributed among all the stations, instead of centralized on a master station, has been pointed out. This feature, which entails the access scheme being partially redesigned, seems to be attractive due to the significant reduction in the queuing delay during traffic transients. A full investigation of distributed algorithms and a comparison with centralized algorithms are made in referenced papers.

Acknowledgments

The author wishes to thank Renzo Beltrame, Erina Ferro and Francesco Potortì for their invaluable help.

References

[1] N. Celandroni and E. Ferro, The FODA-TDMA satellite access scheme: presentation, study of the system and results, *IEEE Trans. Comm.,* **39** (1991) 1823 - 1831.

[2] N. Celandroni, E. Ferro, N. James and F. Potorti', FODA/IBEA: a flexible fade countermeasure system in user oriented networks, *International Journal of Satellite Communications,* **10** (1992) 309 - 323.

[3] N. Celandroni, E. Ferro and F. Potorti', Experimental results of a demand-assignment thin route TDMA system, International Journal of Satellite Communications, *International Journal of Satellite Communications,* **14** (1996) 112-126.

[4] N. Celandroni, E. Ferro and F. Potorti', Comparison between distributed and centralized demand assignment TDMA satellite access schemes*, International Journal of Satellite Communications,* **14** (1996) 95-111.

[5] N. Celandroni, F. Potorti' and S. T. Rizzo, Signal Quality Monitoring to Counter Rain Fading with Adaptive Information Bit Energy in Satellite Thin Route TDMA Systems, CNUCE Report C95-44, December 1995.

[6] N. Celandroni, E. Ferro and F. Potorti', Study of distributed algorithms for satellite channel capacity assignment in a mixed traffic and faded environment. Part II. The FEEDERS-TDMA proposal, submitted to the *International Journal on Satellite Communications*.

[7] W. Petr David, K. M. S. Murthy, S. Frost Victor and A. Neir Lyn, Modelling and Simulation of the Resource Allocation Process in a Bandwidth-on-demand Satellite Communication Network, *IEEE Journal on Selected Areas in Communications,* **10** (1992) 465 - 477.

[8] J.-F. Chang and S.-H. Lin, Delay Performance of VSAT-Based Satellite Wide Area Networks, *International Journal of Satellite Communications,* **11** (1993) 1 - 12.

[9] S. Tasaka Shuji and Y. Ishibashi, A Reservation Protocol for Satellite Packet Communication. - A Performance Analysis and Stability Considerations, *IEEE Trans. Comm.,* **32** (1984) 920 - 927.

[10] F. Delli Priscoli, M. Listanti, A. Roveri and A. Vernucci, A Distributed Access Protocol For An ATM User Oriented Satellite System, *Proc. ICC*, Boston, USA, 1989.

[11] A. Baiocchi, M. Carosi, M. Listanti and A. Roveri, Modelling of a Distributed Access Protocol for an ATM Satellite System: An Algorithmic Approach, *IEEE Journal on Selected Areas in Communications* **9** (1991) 65 - 75.

[12] S. S. Lam, Satellite Packet Communication. Multiple Access protocol and Performance, *IEEE Trans. Comm.,* **28** (1980) 468 - 488.

[13] P. Papantoni-Kazakos and T. Hall, Multiple Access Algorithms For a System With Mixed Traffic: High and Low Priority, *Proc. ICC*, Boston, USA, 1989.

Appendix

The following L-transformed expression of the queue length

$$Q(s) = \frac{[1 - HKe^{-sT}]I(s) - X(s)}{s + Ke^{-sT}}$$

can be written as

$$Q(s) = \frac{1}{s^2}\{(\mathrm{i} - \mathrm{x} - \mathrm{i}e^{-sT})\sum_{j=0}^{\infty}[-1]^j(\frac{K}{s})^j e^{-jsT}\}$$

from which the inverse transform, giving the temporal function is easily obtained

$$q(t) = \sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{(\mathrm{i} - \mathrm{x})[t-(j-1)T]^j u(t-(j-1)T)K^{j-1} - \mathrm{i}HK^j(t-jT)^j u(t-jT)\}.$$

The following L-transformed expression of the system throughput

$$O(s) = e^{-sT/2}C(s) + \frac{[1+sH]Ke^{-s(T+T/2)}I(s) + e^{-sT/2}\ \ s\ \ X(s)}{s + Ke^{-sT}}$$

can be written as

$$O(s) = \frac{\mathrm{c}}{s}e^{-sT/2} + \frac{1}{s^2}\{[\mathrm{i}K(1+sH)e^{-s(T+T/2)}) + \mathrm{x}\ s\ \ e^{-sT/2}]\sum_{j=0}^{\infty}[-1]^j(\frac{K}{s})^j e^{-jsT}\}$$

from which the inverse transform, giving the temporal function is easily obtained

$$o(t) = u(t-\frac{T}{2})\mathrm{c} + \sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{\mathrm{i}K^j(t-jT-\frac{T}{2})^{j-1}[t+(H-T)j-\frac{T}{2}]u(t-jT-\frac{T}{2}) +$$
$$\mathrm{x}\ K^{j-1}j[t-(j-1)T-\frac{T}{2}]^{j-1}u(t-(j-1)T-\frac{T}{2})\}$$

The following L-transformed expression of the total amount of data transmitted and received by the destination station

$$D_O(s) = \frac{O(s)}{s}$$

can be written as

$$D_o(s) = \frac{c}{s^2} e^{-sT/2} + \frac{1}{s^3}\{[\mathrm{i}K(1+sH)e^{-s(T+T/2)}) + \mathrm{x}se^{-sT/2}]\sum_{j=0}^{\infty}[-1]^j(\frac{K}{s})^j e^{-jsT}\}$$

from which the inverse transform, giving the temporal function is easily obtained

$$d_o(t) = u(t-\frac{T}{2})(t-\frac{T}{2}) + \sum_{j=1}^{\infty}\frac{[-1]^{j-1}}{j!}\{\mathrm{i}K^j[\frac{(t-jT-\frac{T}{2})^{j+1}}{j+1} + H((t-jT-\frac{T}{2})^j]u(t-jT-\frac{T}{2})] +$$

$$\mathrm{x}\ K^{j\text{-}1}(t\text{-}(j\text{-}1)T\text{-}\frac{T}{2})^j u(t\text{-}(j\text{-}1)T\text{-}\frac{T}{2})\}$$