
N. Blefari-Melazzi, M. Femminella, G. Reali

Dipartimento di Ingegneria Elettronica e dell’Informazione (DIEI)
University of Perugia, Via G. Duranti, 93 - 06125 Perugia – Italy; E-mail: blefari@diei.unipg.it

Abstract -- Current satellite systems operate according to circuit switching transfer modes. To improve flexibility and efficiency, several kinds of packet switching systems have been proposed [e.g. 7,8]. However, it appears that full packet switches are still considered complex and expensive, when implemented on board the satellites. For the time being, a compromise has been found in satellite networks with Dynamic Bandwidth Allocation Capabilities (DBAC). Such systems are based on classical circuit switches, but the DBAC payload allows changing dynamically the capacity of each connection, without the need of tearing-down and setting-up again the connection itself.

In this paper we consider a DBAC satellite system and define algorithms to allocate the bandwidth so as to provide deterministic and statistical QoS guarantees. The traffic sources are regulated by standard Dual Leaky Buckets (DLBs).

We define bandwidth-handling policies, design Connection Admission Control rules and evaluate analytically the system performance. As expected, the numerical results show a significant increase of the overall utilisation factor of our system, when compared with a plain circuit switching solution.

I. INTRODUCTION

This work is framed in the ACTS program ASSET (ACTS Satellite Switching and End to End Trials), sponsored by the European Union. ASSET is based on the EuroSkyWay (ESW) Network [1,7], which is composed of a number of GEO satellites, a Master Control Station (MCS, on earth), user terminals (mobile), earth stations (fixed), including a certain number of Interworking Units (IWU). These IWUs are stations that interface the ESW system with terrestrial networks (e.g. the Internet, ATM, etc.). In addition to its own protocols, which allow ESW users to exchange information among themselves, the ESW system is designed to support a variety of communication paradigms (e.g. the Internet, the ATM/B-ISDN and the N-ISDN).

It is well known that packet switching provides flexibility and efficiency. However, it appears that full packet switches are still considered complex and expensive, when implemented on board the satellites (at least in the mid-term). In addition, circuit switching has still some attractions, at least for particular services. For the time being, a compromise between circuit and packet transfer modes has been found in satellite networks with Dynamic Bandwidth Allocation Capabilities (DBAC). Such systems are based on classical circuit switches, but the DBAC payload allows changing dynamically the bandwidth of each connection, without the need of tearing-down and setting-up the connection itself. The ESW network is a DBAC system. In ESW, a common framing structure is defined and applied to all the carriers. The smallest unit switched within the system is the so-called satellite cell. Each satellite cell has a size equal to 53 bytes and provides a net information rate of 16000 bit/s (basic channel).

The traffic control methodology is similar to the one of the ATM and to the RSVP paradigms (i.e. Declaration of Traffic descriptors, Connection Admission Control, Policing). The set-up procedure is handled by the MCS; therefore the signalling information has to make two satellite hops (lasting two satellite round trip times) to go from a terminal to the MCS and come back. Once a connection is set-up, the Dynamic Bandwidth Allocation Capabilities allow changing the bandwidth of the connection itself by means of In Band Requests (IBRs). Each user can ask to increase or decrease dynamically (frame by frame) the capacity of a connection, according to his needs. The IBRs are sent to a module resident in the satellite and called Traffic Resource Manager (TRM). The TRM can accept or not the IBRs. When a user requests a change of capacity, the new value of capacity is effective only after that he receives a positive answer from the TRM. The latency of the proposed scheme is then equal to one satellite round trip time, $\Delta T$. Thanks to the DBAC, the ESW remains a fairly simple system (the on-board switch is a simple T-switch) and yet it can offer increased efficiency and flexibility.

In this paper we propose a resource management scheme that guarantees the user perceived performance and exploits the system resources efficiently. In particular, we define bandwidth-handling policies, design Connection Admission Control (CAC) rules and evaluate analytically the system performance. The structure of this paper is as follows. In Section 2 we introduce the system architecture and the proposed resource management scheme. The Connection Admission Control (CAC) rules are defined first with reference to each satellite terminal (in Section 3) and then at the Master Control Station level (in Section 5). Section 4 and 6 are concerned with the relevant performance evaluation. Some concluding remarks are given in Section 7. An Appendix reports material not essential to the continuity of the text.

II. SYSTEM ARCHITECTURE

To simplify the system procedures we make two assumptions:

- the overall traffic is divided into high priority (HP) traffic and low priority (LP) traffic. A given amount of resources is assigned to the HP traffic so as to satisfy pre-defined performance. When the HP traffic is not exploiting all the allocated capacity, the unused capacity is temporarily taken away from the HP traffic and used to carry LP traffic.
When the HP traffic needs the capacity that has been taken away, said capacity is once again given back to the HP traffic itself, after a round trip time. Obviously, these variations of the capacity used by the HP traffic are performed by means of the DBAC.

- the HP traffic entering the system is regulated by means of Dual Leaky Buckets (DLBs). A DLB implements jointly two Generic Cell Rate Algorithms (GCRA) [2,4]. The DLBs assure that the traffic entering the system can be characterised by only four parameters and greatly simplify the three phases of a traffic control scheme, namely the Traffic Descriptors Declaration, the Connection Admission Control (CAC) and the Usage Parameter Control (UPC).

The four parameters controlled by a DLB are: the Peak Cell Rate and its Tolerance, the Sustainable Cell Rate and its Tolerance (or Burst Tolerance). The DLBs can be used also to smooth the incoming traffic.

With reference to the traffic coming out from a generic DLB and feeding the satellite network we denote as:

- $P_T$: the Peak Cell Rate (in cells/s);
- $r_S$: the Sustainable Cell Rate (in cells/s);
- $B_{TS}$: the Token Buffer Size (in cells) (i.e. a parameter related to the Tolerance of the Sustainable Cell Rate, or Burst Tolerance).

We assume that the Tolerance of the Peak Cell Rate is equal to zero or included in the Peak Cell Rate parameter. Thanks to the above assumptions, each HP sources is described by means of three Traffic Descriptors (TDs) (i.e. $P_T$, $r_S$ and $B_{TS}$).

In the following we consider a generic satellite terminal (this can be a mobile terminal, a fixed earth station or an InterWorking Unit, IWU) and we assume that this terminal intends to generate a given amount of traffic towards the satellite network. We assume that each terminal has a buffer dedicated to the HP traffic (HP buffer). We also assume that, if various traffic sources belonging to the same terminal want to emit traffic towards the same destination, then they can be statistically multiplexed over the same ESW connection. This can be an unusual case for a mobile terminal, but it can be a frequent situation for IWUs or for connections carrying traffic on behalf of Service Providers. Note also that a satellite terminal can support different “user terminals” (e.g. a number of N-ISDN telephones), even physically disjoint from the satellite terminal itself and abutted to the latter.

The system architecture of a satellite terminal (SaT) is shown in Fig. 1. For the sake of simplicity we consider only one ESW connection. The HP traffic generated by a SaT and supported by the same ESW connection can be a superposition of $K$ heterogeneous traffic sources. An additional buffer (LP buffer), can be eventually used by the LP traffic. The resource management scheme assures that the performance measures perceived by the HP traffic are always the contracted ones, while no guarantee is given to the LP traffic. The latter could be an IP best effort or an ATM ABR traffic.

We envisage two ways of operation:

Case $a$: deterministic QoS guarantees: at connection set-up, the SaT requests the MCS to establish a connection with a capacity equal to $C$ cells/s. The MCS accepts connection requests until the sum of the relevant capacities is less than or equal to the overall capacity of the involved links (i.e. with a peak allocation). Given the circuit switching transfer mode of the ESW system, this means that the SaT requests a circuit with capacity $C$. However the HP traffic uses transport resources only when they are effectively needed and the capacity actually used is then a function of time, $C(t)\leq C$. The remaining capacity $C - C(t)$ can be used by LP traffic (belonging to the same SaT or to other SaTs). The capacity $C$ is then shared between HP and LP traffic in a dynamic way, as a consequence of IBRs. With this approach the HP traffic does not experience loss phenomena.

Case $b$: statistical QoS guarantees: it works as the above solution with the difference that the MCS does not exert a peak allocation but a statistical one; loss phenomena can occur, but a suitable CAC assures that the performance measures perceived by the HP sources are always the requested ones.
from the HP sources, when not needed, and it takes a round trip time to give it back). The former case is generally called a dimensioning problem while the latter one is referred to as definition of CAC rules. For the sake of simplicity, in the following we will generally speak of “SaT related CAC rules”, being obvious that the same rules can be used both for dimensioning and for resource allocation purposes.

The DBAC has then a twofold application purpose:
1. At MCS level, it improves the system efficiency by allowing the allocation of a number of SaTs greater than the number that would be achieved by means of a peak allocation (in Case b above);
2. Once a SaT is assigned a given connection, the bandwidth may be handled dynamically, without involving the MCS, so that also LP traffic can be delivered, and the channel utilisation factor improves further.

Finally, the possible statistical multiplexing of $K$ sources on the same ESW connection (Fig. 1) helps in increasing the overall efficiency.

The SaT related CAC rules are an essential point of our scheme. In fact, given the system latency $\Delta T$, the HP traffic can not immediately use the capacity that it needs. This means that the values of $C$, $B$ and $K$ must be suitably chosen, as a function of the TDs of the HP traffic sources and of the delay and loss performance requested by the HP traffic. This applies also when a connection supports only one HP source. In fact, a single HP source could experience loss and delay, even if the capacity requested at connection set-up is equal to its Peak Cell Rate, unless a suitable storage capacity is used and a proper CAC is exerted. This happens because the capacity of the connection is not always immediately available, owing to the adopted DBAC scheme and to the system latency. This explains why a SaT must always adhere to CAC rules.

Finally, we remark that we have distributed the CAC rules between MCS and SaTs. A different solution could have been to implement also the SaT related rules in the MCS, thus simplifying the terminals. However, in this case, the “connection request” messages should contain also the TDs of each supported HP traffic source, the performance measures requested by each HP source and the size of the HP buffer of each SaT. This would make more complex the MCS and increase the signalling traffic and the relevant delay.

Thanks to the significant simplification implied by the DLBs, the CAC problem becomes simpler. To define the SaT related CAC rules we start from the ones envisaged in [3 and 10] and we suitably modify them. In [3] the Authors consider a generic network element, composed by a statistical multiplexer with a FIFO queue, fed by traffic sources regulated by means of DLBs. With reference to such network element, they design simple and effective CAC rules that allows guaranteeing Quality of Service (QoS) requirements in loss and delay.

In the following we will define the SaT related CAC rules so as to avoid loss phenomena at the terminal (lossless multiplexing). It is obvious that an alternative solution is to allow small losses at the SaT. This alternative is not discussed here for space limitations. The main difference between our system and the one considered in [3] is that the server of the HP traffic in our system goes on vacation according to a given law (to be described in the following). In other words, the HP server capacity is not constant but it varies dynamically as a consequence of IBRs. To complete the definition of our scheme we must describe how and when a SaT decides to request more capacity to transport HP traffic or to release a portion of the capacity that it is has been using for the HP traffic. In other words, we must define the law followed by the HP server to go on vacation. This will be done in the next section.

### III. SaT RELATED CAC RULES

In this section we assume that the MCS CAC procedure has assigned a given bandwidth, say $C$, to a connection in a given SaT. Eventual modifications of the capacity actually used may happen in every frame, but, when the SaT asks for a bandwidth variation, this variation is effective only when the TRM module on board the satellite acknowledges the request.

The algorithm used to dynamically assign capacity to the HP traffic of a given connection (and, as a consequence, to the LP traffic) is an important issue of our scheme. We assume that the value of $C(t)$, that is the capacity being used by the HP traffic, is chosen as a function of the content of the HP buffer. The remaining capacity, $C-C(t)$, is offered to the LP traffic. It is clear that $C(t)$ must be a non-decreasing function of the HP buffer content: when the content of such buffer increases, the HP traffic requires more capacity and vice-versa. We considered [5] different laws of variation of $C(t)$, as a function of the HP buffer occupancy, $v(t)$.

The simplest of these laws, denoted as $C(v)$, is shown in Fig. 2. When $v(t)$, (i.e. the HP buffer occupancy) is less than a given threshold, denoted as $v_0$, the capacity assigned to the HP traffic is equal to a value denoted as $C_{\text{min}}$; when $v(t)$ is greater or equal to $v_0$ then $C(t)=C$.

For comparison purposes, we define also a Reference system. Such system is a plain circuit switching one, without DBAC capabilities, where the bandwidth assigned to each SaT is constant. Also in such system we foresee a buffer of size $B$, where several HP sources can be statistically multiplexed on the same connection, by using the multiplexing rules defined in [3]. The Reference system is then a system where each ESW connection transports only the HP traffic. The server that models the connection capacity of the Reference system does
not go on vacation and it must be constantly available to transport HP traffic, i.e. $C(t) = C, \forall t$.

The mean value of the capacity used by our system to deliver HP traffic is less than that of the Reference system. If we want the HP traffic to perceive the same loss performance (i.e. zero loss) of the Reference system, it is necessary that a suitable amount of buffer space, $\Delta B$, be dedicated to the compensation of the lower value of the average capacity and of the system latency. Such buffer space, $\Delta B$, may be added to $B$ and, in this case, we can allocate the same number of HP sources as in the Reference system. Otherwise, $\Delta B$ may be seen as a part of the overall buffer size $B$. In the latter case, the buffer that remains to handle the “pure” multiplexing of the HP sources is lower ($B-\Delta B$), therefore less HP sources ($\Delta K$) can be supported than in the Reference system. The parameter $\Delta B$ will be referred to as $\Delta B_{\text{ov}}$, when added to $B$ and as $\Delta B_{\text{in}}$ when considered as a part of $B$.

To evaluate $\Delta B$ we assume that all the HP sources belong to the same class (homogeneous sources) and are characterised by the same values of their TDs ($P_s, r_s$, and $B_{\text{hp}}$). We consider a single ESW connection in a given SaT. The following results can be extended to the heterogeneous sources case.

We establish the following notations:

- $B$: the HP buffer size (in cells);
- $v(t)$: the HP buffer occupancy, at the time $t$ (in cells);
- $C$: the overall capacity allocated to the considered connection (in cells/s);
- $K$: the number of HP sources supported by the considered connection in the considered SaT;
- $C(t)$: the capacity used, at the time $t$, by the HP traffic belonging to the considered connection;
- $\Delta T$: the delay for putting into effect the capacity assignment command, mainly due to the round trip time of the satellite link (in seconds);
- $P_{\text{rot}}$: the total utilisation factor of the considered connection.

Let us now consider a generic function $C(v)$. We assume the token buffer of the shapers to be full at the initial time of a busy period, $t_i$. During the busy period, if $K$ DLB regulated HP sources are emitting traffic at the peak rate $P_s$, the following relation applies:

$$\frac{d(v(t))}{dt} = KP_s - C(v(t) - \Delta T) \quad t_i \leq t \leq t^* = \frac{B_{\text{TS}}}{P_s - r_s}$$

Without loss of generality we may assume $t_i = 0$. The solution of the above differential equation allows determining $\Delta B$. In fact, by integrating (1) in the interval $[0, t^*]$ we obtain:

$$v(t) = KP_s t - \int_0^t C(v(t) - \Delta T) dt$$

Since $KP_s \geq C$, the maximum of $v(t)$ is reached for $t = t^*$:

$$v(t^*) = KP_s t^* - \int_0^{t^*} C(v(t) - \Delta T) dt$$

If all the capacity $C$ were always available to the HP traffic, as in the Reference system, we would obtain:

$$v_c(t^*) = (KP_s - C)t^*$$

Instead, our system needs an additional buffer space, with respect to the Reference system; this additional buffer space is equal to:

$$\Delta B = v_c(t^*) - v(t^*) = C t^* - \int_0^{t^*} C(v(t) - \Delta T) dt$$

The above relation shows that the value of $\Delta B$ depends, by means of an integral relation, on the capacity assignment law, $C(v)$. Therefore we must choose a specific law in order to evaluate $\Delta B$ and to define the CAC rules. However, it can be shown that, by choosing suitably the parameters of one specific law (e.g. the one shown in Fig. 2), it is possible to obtain the same value of $\Delta B$ implied by whatever other law one may choose. In fact, if we are willing to use a compensation buffer space of size equal to a value $K B$, then the only requirement that we have to satisfy is that the mean value of $C(t)$ in the interval $[0, t^*]$ is:

$$\frac{\int_0^{t^*} C(t) dt}{t^*} = \frac{\int_0^{t^*} C(v(t) - \Delta T) dt}{t^*} = C - \frac{\Delta B}{t^*}$$

The condition (6) can be satisfied by an appropriate choice of the parameters of the assumed law of capacity assignment (i.e. $v_c$ and $C_{\text{ov}}$). This result derives from (5) and can be further clarified by Fig. 3, which shows, as a shaded area, a graphical interpretation of the quantity $\Delta B$, for a generic law of capacity assignment.

![Figure 3: Graphic interpretation of the quantity $\Delta B$](image_url)

All this means that, as far as the value of size of $\Delta B$ is concerned, we can restrict ourselves to use always the same capacity assignment law (the one shown in Fig. 2), without loss of generality. We are now interested to see if the same property applies also to other system parameters, such as:

- the number of acceptable HP sources, $K$;
- the maximum delay perceived by the HP users in the HP buffer, $D_{\text{max}}$.

If that were the case, we could avoid considering different laws of capacity assignment. In the following, we will refer to such property as to the Law Independence property, or LI property. We have seen that the LI property applies to $\Delta B$. We now check if the LI property applies also to the number of
acceptable HP sources, \( K \). To this end we have to distinguish if our system and the Reference system have the same buffer size or not. Let us consider the first case. The quantity \( \Delta B \) is a part of the overall buffer size \( B \). A connection with capacity \( C \) can support less HP sources than in the Reference system (\( \Delta K \)) and the parameter \( \Delta B \) will be referred to as \( \Delta B_{IN} \).

Let \( c \) and \( b \) denote the effective bandwidth and buffer allocation, respectively, of the HP sources [3]. In this case:

\[
C = (B - \Delta B_{IN})/b = K
\]  

(7)

Since we want to avoid loss phenomena at the HP buffer, the following relation must hold [3]:

\[
\frac{P_s - c}{P_s - r_s} B_{TS} = b
\]  

(8)

From (7) and (8) we obtain:

\[
(KP_s - C)t^i = B - \Delta B_{IN}
\]  

(9)

Equation (9) shows the dependence of \( \Delta B_{IN} \) on \( K \) and proves that the LI property applies also to \( K \), since it applies to \( \Delta B_{IN} \) when the compared systems have the same overall buffer size. When \( \Delta B \) is added to \( B \) (i.e. \( \Delta B = \Delta B_{IN} \)) both systems support the same number of HP sources, \( K \). In this case, it is evident that \( K \) does not depend on the capacity assignment law. Therefore we can conclude that the LI property applies to \( K \).

The LI property does not apply, in general, to the maximum delay, \( D_{max} \). However, if a given requirement on \( D_{max} \) can be satisfied by adopting the chosen law, then, given its simplicity and the fact that the LI property applies to all the remaining performance measures of interest, we can always adopt such law, without loss of generality. On the other side, we verified that this law can satisfy a wide range of values of \( D_{max} \). For these reasons, in the remainder of this paper, we consider only the above capacity assignment law. Thanks to such assumption we can first particularise the relations given above and then design the SaT related CAC rules. By expressing the above equations as a function of the parameters of the assumed law of capacity assignment (i.e. \( v_i \) and \( C_{min} \)) we get:

\[
\begin{align*}
\nu(t) &= (KP_s - C_{min})t^i, \quad 0 \leq t \leq t_0 + \Delta T \\
\nu(t) &= (KP_s - C_{min})(t_0 + \Delta T) + \\
&\quad + (KP_s - C)(t - t_0 - \Delta T), \quad t_0 + \Delta T < t \leq t^i
\end{align*}
\]  

(10)

where \( t_0 = v_0/(KP_s - C_{min}) \) is the first time when the threshold \( v_0 \) is reached. By combining (5) and (10) we have:

\[
\Delta B = \nu(t^i) - v_c(t^i) = (C - C_{min})(t_0 + \Delta T) = \\
(C - C_{min})\left(\frac{v_0}{KP_s - C_{min}} + \Delta T\right)
\]  

(11)

Fig. 4 shows a graphic representation of the compensation buffer size, \( \Delta B \); in this figure the curves \( v_i(t) \) and \( v_c(t) \) are given by:

\[
v_i(t) = (KP_s - C), \quad 0 \leq t \leq t^i
\]  

(12)

With these results, we can derive the SaT related rules, starting from the case of \( \Delta B = \Delta B_{IN} \). To this aim, we must determine the unknown quantities \( c, b, K, \) and \( \Delta B_{IN} \), by means of the following system of equations (which collects (7), (8), and (11)):

\[
\begin{align*}
\Delta B_{IN} &= (C - C_{min})\left(\frac{v_0}{KP_s - C_{min}} + \Delta T\right) \\
\frac{P_s - c}{P_s - r_s} B_{TS} &= b \\
\frac{C}{c} &= K \\
\frac{C}{b} &= \Delta B_{IN}
\end{align*}
\]  

(14)

This non-linear system is relatively easy to solve. However, we must check the physical acceptability of the mathematical solutions provided by (14). The solutions are physically acceptable if \( \Delta B < B \) and \( r_s \leq P_s \). We also assume that \( r_s < 1 \).

Following the approach of [3], the solution of the above system may be graphically represented as shown in Fig. 5. The following considerations are in order.

- The maximum number of HP sources that can be accepted, \( K \), is a number belonging to a range \( I \) contained in the interval \( [C/P_s, C/r_s] \); such range \( I \) is a function of the system parameters and decreases when \( \Delta B_{IN} \) increases.
- When the threshold \( v_0 \) increases and \( C_{min} \) decreases, the value of \( \Delta B_{IN} \) increases, the effective bandwidth tends to \( P_s \) and the corresponding effective buffer vanishes. This means that \( B = \Delta B_{IN} \). The entire HP buffer is used to compensate the lower value of the mean value of the capacity usable by the HP traffic and the system latency. There is no buffer available for multiplexing purposes. This also implies that the capacity \( C_{min} \) can not be lower than a
suitable bound, which depends on the value of the threshold \( v_o \), otherwise we would obtain unacceptable solutions from the system of equations (14), that is

\[
C_{\text{min}} \geq C - (B - v_0) / \Delta T
\]

(15)

\[
\Delta B_{\text{vs}} = \frac{B - v_0}{c}
\]

and \( B \). To determine the additional relation that gives the desired value of \( B \), we start by recognising that the maximum delay, \( D_{\text{max}} \), is equal to the delay suffered by the last cell in a full HP buffer.

This delay is composed of two contributions:

1. a period lasting \((B - v_0) / C + \Delta T\), during which the buffer is emptied at a rate \( C \);
2. a period lasting \((v_0 - C\Delta T) / C_{\text{min}}\), during which the buffer is emptied at a rate \( C_{\text{w0}} \).

The additional relation is then:

\[
(B - v_0) / C + \Delta T + (v_0 - C\Delta T) / C_{\text{min}} = D_{\text{max}}
\]

Finally, we remark that, if \( \Delta T \geq v_0 / C \), the maximum delay is equal to that of the Reference system (and the second contribution above is equal to zero).

IV. SaT RELATED CAC RULES PERFORMANCE

A comprehensive presentation of the performance of the SaT related CAC rules can be found in [5]. In this section we present only the main results, by focusing on the capacity made available to carry LP traffic and on the total utilisation factor of the considered connection, \( \rho_{\text{tot}} \) (i.e. the utilisation deriving from the superposition of HP and LP traffic). It is worth stressing that we evaluate the capacity that the LP traffic can potentially use and not the one effectively used. The difference between these two quantities depends on the LP traffic generation characteristics and on the LP buffer size.

The capacity made available to carry LP traffic depends on the output processes of the DLB regulators and these, in turn, depend on their input processes, that is to say on the statistical properties of the sources loading the DLBs. This means that each particular kind of source feeding the DLB gives rise to different performance measures of our system. Instead of evaluating the system performance by assuming specific kinds of sources, we will determine upper and lower bounds of the above performance measures. Such bounds do not depend on the particular sources feeding the DLBs.

We consider two extremal behaviours of the output process of a generic DLB. The first one, is a DLB that emits cells at a constant rate equal to the Sustainable Rate, \( r_s \). The second one is a DLB that emits cells according to an ON-OFF process; in the ON state, it transmits cells at a rate \( P_s \); in the OFF state it does not transmit. The length of the ON and OFF periods are such that the overall mean rate is equal to \( r_s \). We conjecture that the first behaviour results in the best possible performance, as far as the capacity made available to carry LP traffic is concerned, while the second behaviour gives rise to the worst possible performance. Accordingly, the parameters relevant to these two behaviours will be labelled with BEST and WORST, respectively. In both these cases we will evaluate \( \rho_{\text{tot}} \) as a function of the parameters of the assumed capacity assignment law \((v_o \text{ and } C_{\text{w0}})\).

The capacity made available to carry LP traffic depends on the output processes of the DLB regulators and these, in turn, depend on their input processes, that is to say on the statistical properties of the sources loading the DLBs. This means that each particular kind of source feeding the DLB gives rise to different performance measures of our system. Instead of evaluating the system performance by assuming specific kinds of sources, we will determine upper and lower bounds of the above performance measures. Such bounds do not depend on the particular sources feeding the DLBs.

The capacity made available to carry LP traffic depends on the output processes of the DLB regulators and these, in turn, depend on their input processes, that is to say on the statistical properties of the sources loading the DLBs. This means that each particular kind of source feeding the DLB gives rise to different performance measures of our system. Instead of evaluating the system performance by assuming specific kinds of sources, we will determine upper and lower bounds of the above performance measures. Such bounds do not depend on the particular sources feeding the DLBs.

Figure 5: Determination of effective bandwidth and effective buffer for lossless multiplexing

Let us now consider the case \( \Delta B = \Delta B_{\text{out}} \). The total buffer size is

\[
B_{\text{TOT}} = B + \Delta B_{\text{out}}
\]

In this case, we can allocate the same number of HP sources, \( K \), as in the Reference system. The number \( K \) can be obtained by solving the system of equations:

\[
\begin{align*}
\frac{P_s - c}{P_s} B_{\text{TS}} &= b \\
\frac{c}{c} &= \frac{B}{b} \\
\frac{c}{c} &= K
\end{align*}
\]

(16)

The corresponding overall buffer size is given by:

\[
B_{\text{TOT}} = B + (C - C_{\text{min}}) \left( \frac{v_0}{K P_s - C_{\text{min}}} + \Delta T \right)
\]

(17)

The above rules guarantee that our system and the Reference one have the same loss performance (zero loss). If we want that also the maximum delay suffered by the HP traffic be equal in both system we must add a further constraint. We conclude then this Section with the definition of rules including such constraint. Since the maximum delay must be not greater than that suffered in the Reference system, the compensation buffer space, \( \Delta B \), must be included in \( B \), therefore the parameter \( \Delta B \) will be referred to as \( \Delta B_{\text{vs}} \). In addition, the average capacity used by the HP traffic is smaller than that of the Reference system and less than \( C \). This means that the size of the overall buffer size of our system must be smaller than that of the Reference system. The value of \( B \) must then be determined by using a suitable constraint, which must be added to those of (14). The set of unknown quantities of the resulting system of equations is therefore given by \( c, b, K \), \( \Delta B_{\text{vs}} \) and \( B \).
\begin{align}
\rho_{TOT,WORST} &= 1 - \frac{P_s - r_s}{P_s} \left( 1 - \frac{r_s D_{MAX}}{B_{TS}^2} \right) \frac{C_{min}}{C} \tag{19b} \\
\text{for } \Delta T &\leq v_0 / C \ , \text{ while} \\
\rho_{TOT,WORST} &= 1 - \frac{r_s (P_s - r_s)}{B_{TS}^2 P_s} \left( \Delta T - v_0 / C - \frac{1}{r_s} \frac{C_{min}}{C} \right) \tag{19c} \\
\text{for } \Delta T &> v_0 / C \ ,
\end{align}

where \( D_{MAX} \) is given by (18) and

\[
\gamma = \frac{B_{TS}}{P_s - r_s} \left( \frac{P_s}{r_s} - 1 \right) - \frac{B - v_0}{C} - \Delta T
\]

In the WORST case, the system is stable only if:

\[
C_{min} \geq \frac{v_0 - C \gamma}{P_s - r_s} \left( \frac{P_s}{r_s} - 1 \right) - \Delta T - \frac{B - v_0}{C}
\]

(20)

Fig. 6 shows the upper (BEST) and lower (WORST) bounds of \( \rho_{tot} \) as a function of \( C_{min}/C \) when \( \Delta B = \Delta B_{opt} \); the loss performance and maximum delay are identical to those of the Reference system (Eqs. 14 and 18). For comparing the proposed approach to the Reference system, a solid line corresponding to the utilisation factor of the Reference system is plotted, and labelled “(no DBA)”. The values of the system parameters are:

- \( P_s = 32000/\text{(48 x 8)} = 83.33 \text{ cells/s} \);
- \( r_s = 32.197 \text{ cells/s} \);
- \( B_{TS} = 100 \text{ cells} \);
- \( C = 2048000 / (53 \times 8) = 4830.189 \text{ cells/s} \);
- \( \Delta T = 0.25 \text{ s} \).

![Figure 6: Overall utilisation factor of the considered connection, as a function of \( C_{min}/C \)](image)

As the ratio \( C_{min}/C \) varies, so does the number of HP sources that can be allocated in our system \((K)\) and the value of the buffer size \((B)\) needed to avoid loss phenomena in our system. The Reference system can support 76 sources \((K_{ref, opt} = 76)\) and its buffer size is equal to 3000 cells \((B_{0.05, opt} = 3000)\). A dimensioning criterion that resulted from the analysis is that an optimal working point corresponds to selecting a value of \( C_{min} \) equal to \( Kr \); in fact, this choice produces a full utilisation factor and the maximum number of HP sources allocated. Should this be not possible, due to boundaries (15) or (20), the value of \( C_{min} \) should be the closest to \( Kr \) that satisfies such constraints. In Fig. 6, the optimal working point corresponds to \( C_{out}/C \) equal to about 0.42.

We can observe that:

- the proposed approach results in a marked increase of the overall utilisation, with respect to the Reference system; this improvement increases when the traffic burstiness (i.e. the ratio \( P/r_s \) increases);
- the price to be paid for such advantage is a reduction of the HP sources, which is compensated by the possibility to transport LP traffic. Note also that our approach uses a smaller buffer size (so as to provide the same maximum delay of the Reference system); if the HP traffic could tolerate a higher delay we could support the same number of HP sources while increasing the overall utilisation factor (case of \( \Delta B = \Delta B_{out} \));
- the capacity made available to the LP traffic tends to decrease when \( C_{min} \) increases, while the number \( K \) of allowed HP sources increases;
- when the value of \( C_{min} \) approaches \( C \), the number of HP sources tends to that of the Reference system, since the DBAC mechanism tends to have no effect.

Other results, including those relevant to the case \( \Delta B = \Delta B_{out} \) and to single sources, can be found in [5].

V. MCS RELATED CAC RULES

As said in Sec. II, the MCS related CAC rules can be defined according to two approaches: peak allocation (and no cell loss) or statistical one with small loss probabilities allowed. The first one is trivial: the maximum number of SaTs that can access simultaneously a link with capacity \( C_T \) is:

\[
H_{NO\ Loss} = C_T / C
\]

(21)

In this case, the only way to increase the link utilisation, with respect to the Reference system, consists in using the DBAC, which allow transporting also LP traffic.

Let us now consider the second approach (statistical allocation). We first assume that all the SaTs have the same characteristics (homogeneous terminals) and then we will extend the results to the heterogeneous case. In any case, we assume that the traffic loading the generic SaT is independent by that of the others terminals.

A. Homogeneous terminals

In order to derive the CAC rules a first (conservative) solution is to assume a worst case behaviour of the output process of the SaT: each SaT alternates between the two emission rates \( C_{min} \) and \( C \) (Fig. 7).

In a time period \( T = P_s f^* / r_s \), the maximum capacity \( C \) is requested for a time \( T_x \) given by:

\[
T_x = \frac{B}{C} - v_0 \left( \frac{1}{KP-C_{min}} + \frac{1}{C} \right)
\]

(22)

During the period \( T-T_x \), the capacity used is \( C_{min} \). We
now assume that the phase displacements of the above emission processes, relevant to all the involved terminals, are uniformly distributed. The assumption to consider all the emission processes synchronised would be too pejorative. We now let:
- \( C_i(t) = C(t - \vartheta_i) \): the instantaneous capacity requested by the \( i \)-th SaT, where \( \vartheta_i \) is uniformly distributed in \([0, T]\);
- \( U(t) \): the total instantaneous capacity requested by all the SaTs;
- \( P_{loss} \): the loss probability suffered by the HP traffic;
- \( H \): the number of SaTs that can be accepted by the MCS in a link with a total capacity equal to \( C_T \), under the requirement that \( P_{loss} \) be less than a given threshold, \( L \).

\[
M(s) = E[e^{s\hat{C}_i}] = (1 - \omega)e^{sC_{min}} + \omega e^{sC_T}.
\]  

We assume that the stability condition \([3]\) is verified:

\[
\sum_{i=1}^{H} E[\hat{C}_i] < C_T
\]  

and that loss phenomena can occur:

\[
\lim_{s \to \infty} H \frac{M'(s)}{M(s)} > C_T
\]  

Figure 7: Worst case capacity request from a single SaT

The value of \( H \) must satisfy the following inequality:

\[
P_{loss} = Pr(U(t) > C_T) = Pr \left( \sum_{i=1}^{H} C(t - \vartheta_i) > C_T \right) = \Pr \left( \sum_{i=1}^{H} C_i(t) > C_T \right) \leq L
\]  

(23)

The last relation in (23) is maximised \([3]\) by the probability that the sum of the \( H \) random variables \( \hat{C}_i \) (where \( \hat{C}_i \) is the random variable relevant to \( C_i(t) \)) exceeds the value \( C_T \); the p.d.f. of \( \hat{C}_i \) is:

\[
p(\hat{C}_i) = \left( 1 - \frac{T_x}{T} \right) \delta(\hat{C}_i - C_{min}) + \frac{T_x}{T} \delta(\hat{C}_i - C)
\]  

(24)

where \( \delta \) is the Dirac function.

An alternative to the worst case behaviour of Fig. 7, which results in a less conservative allocation, is to assume that whenever the HP buffer exceeds the threshold \( \nu_0 \), the SaT emits at the rate \( C \) for a time period equal to \( T_x \) (22), as in the worst case. In this case the probability that the threshold \( \nu_0 \) is exceeded is taken into consideration and (24) becomes:

\[
p(\hat{C}_i) = \omega \delta(\hat{C}_i - C) + (1 - \omega) \delta(\hat{C}_i - C_{min}),
\]  

(25)

where \( \omega = \frac{T_x}{T} \Pr(\nu(t) > \nu_0) \) and \( \Pr(\nu(t) > \nu_0) \) is derived in the Appendix.

To satisfy the bound in relation (23), we use the sum of the random variables \( \hat{C}_i \) instead of the stochastic process \( U(t) \) and we evaluate an upper limit of \( P_{loss} \) by means of the Chernoff’s bound. According to this approach, we must define the moment generating functions for \( \hat{C}_i \), given by:

\[
M(s) = E[e^{s\hat{C}_i}] = (1 - \omega)e^{sC_{min}} + \omega e^{sC_T}.
\]  

We can now use the large deviation approximation to Chernoff’s bound, where \( P_{loss} \) is approximated by the following expression \([3]\):

\[
P_{loss} \leq e^{-sC_T} E[e^{s \sum_{i=1}^{H} \hat{C}_i}] = e^{-sC_T} M(s)^H
\]  

We can make this bound tighter by minimising the function in the variable \( s \) on the right-hand side; this can be done by maximising the function \( F(s) \) defined as \([3]\):

\[
F(s) = sC_T - H \log(M(s)) \]

(31)

This function has a single maximum \( s^* > 0 \), given by:

\[
s^* = -\frac{1}{(C - C_{min})} \log\left( \frac{\omega HC - C_T}{1 - \omega C_T - HC_{min}} \right)
\]  

(32)

We can now use the large deviation approximation to Chernoff’s bound, where \( P_{loss} \) is approximated by the following expression \([3]\):

\[
P_{loss} = e^{-F(s^*)}
\]  

(33)

and \( \sigma^2(s) \) is \([3]\):

\[
\sigma^2(s) = H \left[ \frac{M''(s)}{M(s)} - \left( \frac{M'(s)}{M(s)} \right)^2 \right]
\]  

(34)

Finally, \( H \) can be found by solving the following equation:

\[
\frac{e^{-F(s^*)}}{\sigma^2(s)} = L
\]  

(35)

The effective bandwidth for each SaT is therefore given by:

\[
\bar{C} = C_T / H
\]  

(36)

where \( \bar{C} \) is bounded by \( C_{min} + \omega(C - C_{min}) \) and \( C \).
B. Heterogeneous terminals

This case is deeply discussed in its general form in [3]. It is necessary to consider the superposition of a plurality of stochastic process \( U_j(t) \) that represents the bandwidth demand of each class of terminals \( (j = 1,...N) \). The treatment is analogous to that of the homogenous case, with the presence of different kind of random variables \( \tilde{C}_{ji} \), used to represent the different bandwidth demand of the \( N \) types of terminals. If the maximum cell loss probability tolerated is \( L \), the solution of the problem is the set \( H = \{ H_j, j = 1,...N \} \) that satisfies the following inequality in the tightest way:

\[
R_{\text{loss}} = \Pr \left( \sum_{j=1}^{N} U_j(t) > C_T \right) = \Pr \left( \sum_{j=1}^{N} H_j C_{ji}(t) > C_T \right) \leq L. \tag{37}
\]

Also in this case we can use the Chernoff's bound with the following formalism:

\[
p(\tilde{C}_{ji}) = \left(1 - \omega^j\delta(\tilde{C}_{ji} - C^j_{\text{min}}) + \omega^j\delta(\tilde{C}_{ji} - C^j)\right) \tag{38}
\]

where \( \omega^j = \frac{T_j}{T} \Pr \left( \gamma^j(t) > \gamma^j_0 \right) \) and the derivation can continue as in the case of homogeneous terminals [3,5].

VI. NUMERICAL RESULTS

In this section we present some results with the aim of validating the approximation of the threshold crossing probability presented in the Appendix and of evaluating the performance of the proposed approach. The values of the system parameters are equal to those of Sec. IV, with the exception of the total link capacity, \( C_T \), which is now equal to 45 Mbit/s. Fig. 8 shows the analytical upper bound and some simulated values of the threshold crossing probability, \( \Pr(\gamma(t) > \gamma_0) \), as a function of the HP buffer size and for different values of the ratio \( \gamma_0 / B \). The value of \( C^j_{\text{min}} \) is the smallest possible value greater than the optimal working point \( K_r \), note that the bandwidth in our system is quantised with steps equal to the basic channel. The figure confirms that the derivation presented in the Appendix gives an upper bound of \( \Pr(\gamma(t) > \gamma_0) \) and allows evaluating its tightness.

Fig. 9 shows the number of SaTs that can be allocated as a function of the required loss probability, \( L \), and for different values of \( B \) and of the ratio \( \gamma_0 / B \).

We observe that:

1. when \( \gamma_0 / B \) increases, the threshold crossing probability decreases and, for a given buffer value, the average capacity demand from each SaT is sensibly lower;
2. when \( B \) increases, the dependence of \( H \) from the loss probability becomes less evident; this phenomenon is due to the lower values of threshold crossing probability, which tends to make the increase of \( H \) slower. In fact, we must remember that the average bandwidth demand is limited by \( C \) and \( C_{\text{min}} + \omega(C - C_{\text{min}}) \), - see comment after (36) - so that when \( \omega \) is very small, the capacity allocation is always close to \( C_{\text{min}} \).

3. the number of SaTs, \( H \), with lower buffer space \( B=2000 \) can be higher than the one with more buffer space \( B=5000 \) (for \( \gamma_0 = 0.5B \) in this figure). This is because each SaT may support a variable number of traffic sources and it may happen that, when we have a lower number of SaTs we also have a higher number of traffic sources supported by such SaTs (and viceversa). However, the overall supported HP traffic (i.e. the number of HP sources, \( K \)) is always a non-decreasing function of the buffer size (as it must be). In any case, a marked increase of \( H \), with respect to the value relevant to the peak allocation \( H_{\text{NO LOSS}} \), is always evident.
full when a capacity request is denied by the TRM and so the effective cell loss suffered by that SaT is less than the estimated one. This is because $P_{loss}$ has been evaluated as the fraction of time in which there is not sufficient link capacity. An evaluation of the overall efficiency factor resulting from the joint application of SaT and MCS (statistical) related CAC rules can be found in [5].

VII. CONCLUSIONS

In this work we have shown that it is possible to improve significantly the efficiency of a circuit switched satellite system by means of the DBAC and in particular that it is possible:

- to use efficiently the bandwidth assigned to each SaT by releasing a portion of it to the LP traffic;
- to allocate a number of SaTs greater than that relevant to a peak allocation. In this way, even if the bandwidth available for the LP traffic is reduced, the number of HP users increases considerably;
- to increase the connection carrying capacity for given buffer spaces by allowing small cell losses in the HP buffers; this result is not shown here for space limitations;
- to provide QoS guarantees for the HP traffic.

In future work we propose to enhance our approach in several respects. First, we intend to study the performance perceived by the LP traffic, in terms of loss and delay. Second, we want to integrate in our framework a reactive control scheme, to guarantee the requirements in loss of the LP traffic [e.g. 8]. Third we want to improve the approximation in the evaluation presented in the Appendix.

APPENDIX

An upper bound of the $Pr(v(t) > v_0)$ can be evaluated by solving a queuing system that models an HP buffer loaded by a superposition of $K$ WORST case DLBs output processes (that is, deterministic On-Off sources, with an On period consisting of a burst of fixed length equal to $a=t^n P_2$ cells) with random relative phasing. The service time alternates between the values relevant to the output rates $C$ and $C_{min}$. To solve this system we make some approximations: i) we assume an infinite buffer size (which is a conservative assumption, in our case); ii) we assume that the output rate is always constant and equal to $C_{min}$; this allows avoiding the complexity resulting by a service time that is a function of the buffer content. Note that this implies that the value of $C_{min}$ can not be smaller than $Kr$, so that the queue utilisation factor is less than $1$; (note also that, for values of $C_{max}$ near to the optimal working point $Kr$, the Chernoff's bound approach gives for $Pr(v(t) > v_0)$ values very near to one); iii) we approximate the arrivals of all the cells of a burst with an instantaneous aggregated arrival of equivalent size. The latter assumption is made since the exact solution of the queuing system up to now defined and with the above two assumptions (available e.g. in [9]) has a high computational cost. We then normalise all the variables with respect to the burst size $a$, so that we can apply the results obtained in [6] relevant to a $K \cdot D/D/1$ queue, that is a queuing model where each of the $K$ sources generates just one cell in every period $D$ with $D$ (period normalised to the service time) $\bar{v}(t)$ (buffer level) and $\bar{v}_0$ (threshold) defined as:

$$D = \left[ \frac{TC_{min}}{a} \right]$$

$$\bar{v}(t) = \frac{v(t)}{a}$$

$$\bar{v}_0 = \frac{v_0}{a}$$

(39)

Under these hypotheses, the threshold crossing probability becomes [6]:

$$Pr(\bar{v}(t) > \bar{v}_0) = Pr(\bar{v}(t) > v_0) =$$

$$\sum_{s=1}^{K-\bar{v}_0} \left( \frac{K}{\bar{v}_0 + s} \right) \left( \frac{s}{D} \right) \frac{1 - \frac{s}{D}}{D - s} \frac{D - K + \bar{v}_0}{D - s}$$

(40)

REFERENCES