Allocation of QoS Connections in MF-TDMA Satellite Systems: A Two-Phase Approach

Jung-Min Park, *Member*, *IEEE*, Uday Savagaonkar, Edwin K. P. Chong, *Fellow*, *IEEE*, Howard Jay Siegel, *Fellow*, *IEEE*, and Steven D. Jones, *Member*, *IEEE*

***Abstract*—We address the problem of providing guaranteed quality of service (QoS) connections over a multi-frequency time division multiple access (MF-TDMA) system that employs differential phase-shift keying (DPSK) with various modulation modes. The problem can be divided into two parts—resource calculation and resource allocation. We present algorithms for performing these two tasks, and evaluate their performance in the case of a Milstar Extremely High Frequency Satellite Communication (EHF-SATCOM) system.**

In the resource-calculation phase, we calculate the minimum number of timeslots required to provide the desired level of bit error rate (BER) and data rate. The BER is directly affected by the disturbance in the link parameters. We use a Markov modeling technique to predict the worst-case disturbance over the connection duration. The Markov model is trained offline to generate a transition probability matrix, which is then used for predicting the worst-case disturbance level. We provide simulation results to demonstrate that our scheme outperforms the scheme currently implemented in the EHF-SATCOM system.

The resource-allocation phase addresses the problem of allocating actual timeslots in the MF-TDMA channel structure (MTCS). If we view the MTCS as a collection of bins, then the allocation of the timeslots can be considered as a variant of the dynamic bin-packing problem. Because the bin-packing problem is known to be NP-complete, obtaining an optimal packing scheme requires a prohibitive amount of computation. We propose a novel packing heuristic called Reserve Channel with Priority (RCP) fit, and show that it outperforms two common bin-packing heuristics.

Manuscript received November 21, 2002; revised December 30, 2003. A preliminary version of portions of this material was presented in [13]. This research was supported in part by the DARPA/ITO AICE Program under contract numbers DABT63-99-C-0010 and 0012, by NSF under grants ANI-0207892, ANI-0099137, and ECS-0098089, and by the Colorado State University George T. Abell Endowment.

J.-M. Park is with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (phone: 540-231-8392, fax: 540-231-8292, e-mail: jungmin@vt.edu).

U. Savagaonkar is with the Intel Communications Technologies Lab in Hillsboro, OR, USA (e-mail: uday.r.savagaonkar@intel.com).

E. K. P. Chong is with the Department of Electrical and Computer Engineering and Department of Mathematics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: echong@colostate.edu).

H. J. Siegel is with the Department of Electrical and Computer Engineering and Department of Computer Science, Colorado State University, Fort Collins, CO 80523 USA (e-mail: hj@colostate.edu).

S. D. Jones is with the Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723 (e-mail: Steven.Jones@jhuapl.edu).

***Index Terms*—Satellite, resource allocation, QoS, MF-TDMA, bin packing, Markov modeling, prediction.**

I. Introduction

THE multi-frequency time division multiple access (MF-TDMA) scheme is a hybrid solution that combines the strengths of the frequency division multiple access (FDMA) and time division multiple access (TDMA) techniques, and hence is favored by many modern satellite communication systems. This technique allows for efficient streaming of traffic while maintaining flexibility in capacity allocation. Access to the satellite uplink employing this technique is characterized by a large number of connections that share limited system resources. In systems employing MF-TDMA as their uplink access method, multiple frequency channels are allocated for the uplink access, and the TDMA scheme is employed in each frequency channel. Thus, each frequency channel is divided into several timeslots that can be assigned to multiple connections. We treat the timeslots as the resource that needs to be allocated to each connection. Each connection is assigned a fixed portion of the resource based on its quality of service (QoS) requirements. Specifically, we consider two QoS measures—data rate and maximum-allowable bit error rate (BER). It is assumed that each connection declares its QoS requirements at the time of the connection request. We treat the data rate as a deterministic QoS measure, and the BER as a statistical QoS measure—throughout the duration of a connection, a fixed data rate is guaranteed, whereas the maximum-allowable BER is assured with a certain probability. Our aim is to provide QoS guarantees to every connection throughout its duration. To achieve this objective, we concentrate on two specific problems that are limited to the uplink of the MF-TDMA satellite systems—resource calculation and resource allocation. Specifically, we focus on the above two problems applied to the Milstar Extremely High Frequency Satellite Communication (EHF-SATCOM) system. This satellite system is designed to provide reliable communications for the US Military's strategic and tactical forces. (See Section II-A for more details.)

It is impractical to reconfigure a connection once it is allocated a position on the MF-TDMA channel structure (MTCS). A typical reconfiguration of the MTCS for the Milstar EHF-SATCOM system could take as long as 40 seconds or longer. This is a considerable delay relative to the

average connection duration for the system. Hence, reconfiguration of the MTCS, and consequently that of a connection, is undesirable. At the same time, the number of timeslots allocated to a connection directly affects the QoS of the link. To elaborate, in the case of systems employing multiple modulation modes with an MF-TDMA channel structure, the two QoS measures under consideration are directly related to the number of timeslots allocated, the modulation mode being used, and the disturbance level in the system via the link-budget equations. The disturbance level in turn depends on various system and environmental parameters such as transmitter power and rain rate. We will describe these relations briefly in Section II-B.

While some of the parameters contributing to the disturbance level are deterministic, others are not. The aggregated effect of the nondeterministic parameters changes the minimum number of timeslots needed to guarantee the QoS level of a connection during its duration. We present a *Markov model-based prediction* (*MMP*) scheme for predicting the worst-case disturbance level over the connection duration. We use this prediction to compute the number of timeslots required in the worst case. Açar and Rosenberg [1] also have investigated the problem of resource calculation, but their study considered Asynchronous Transfer Mode (ATM) over MF-TDMA satellite links, and they used performance measures different from ours.

After the resource-calculation algorithm determines the number of timeslots required to satisfy the QoS requirements of a connection, a resource-allocation algorithm is needed to map the timeslots onto the MTCS. If we view the frequency channels of the MTCS as a collection of bins, then the problem of allocating resources for the uplink can be viewed as a variant of the *dynamic bin-packing* problem. Motivated by potential applications, such as computer storage, the classical bin-packing problem has been actively researched and analyzed (e.g., [2], [9]). The objective of the classical bin-packing problem is to pack the bins with the given items as densely as possible (i.e., pack the items into as few bins as possible). Because the bin-packing problem is NP-complete [9], most of the research has concentrated on finding upper and lower bounds on the worst-case performance of well-known simple algorithms (e.g., first fit and best fit), rather than searching for an optimal solution. Although these well-known packing algorithms obtain relatively "good" placements for the classical bin-packing problem, the packing restrictions that are unique to the resource-allocation problem of the MTCS make the straightforward application of these algorithms to our problem ineffective.

We propose a novel packing algorithm, called *Reserve Channel with Priority* (*RCP*) *fit*, for the resource allocation in an MTCS. To measure the performance of bin-packing algorithms, one might want to obtain the expected performance of such algorithms under various probabilistic assumptions, such as arrival times, departure times, and size of the items. However, it has been shown that such results are extremely difficult to obtain theoretically, even for static bin packing. Furthermore, even in static bin packing, obtaining numerical indicators for a relatively sophisticated packing procedure under probabilistic assumptions is nearly impossible due to the enormous complexity of the calculations [2]. Thus, we compare the performance of RCP fit with other packing algorithms (i.e., best fit and first fit) via simulations.

In the next section, we introduce the Milstar EHF-SATCOM system and its MF-TDMA uplink channel structure, which was used as the model for the simulation experiments. The resource-calculation and resource-allocation phases are described in Section III and Section IV, respectively. We provide the simulation results in Section V. Finally, in Section VI, we conclude the paper with a discussion of the results. In the Appendix, we prove the NP-completeness of the resource-allocation problem.

II. The EHF-SATCOM System

A. Channel Structure

We adopt a satellite system model based on the Milstar EHF-SATCOM system. This satellite system is designed to provide reliable communications for the US military's strategic and tactical forces. Concepts for survivability in a hostile space environment have shaped the design of this system—it is robust against both electronic warfare and physical attacks carried out by the enemy. The Milstar system is a joint satellite communications system that is designed to provide secure worldwide communications for high-priority military users (i.e., command authorities). The multi-satellite constellation is capable of linking the command authorities with a wide range of military resources (e.g., ships, submarines, and aircraft).

Unlike systems using lower frequencies, Milstar satellite systems utilizing EHF technology (30 ~ 300 GHz) offer numerous advantages[1]:

- able to avoid interference and crowding, which is problematic in other frequency bands;
- rapid recovery from the scintillation caused by a high-altitude nuclear detonation;
- minimal susceptibility to enemy jamming and eavesdropping;
- ability to achieve smaller secure beams with modest-sized antennas.

The EHF-SATCOM system is comprised of three distinct parts—the space segment, the user segment, and the control segment. The satellites correspond to the space segment, earth terminals correspond to the user segment, and the control segment consists of satellite control and planning elements. The system can support multiple voice and data channels originating from many terminals simultaneously. The space segment (satellite) acts essentially as a relay and router in the sky. It receives, demodulates, routes, and re-modulates information flows. The user segment (terminal) is capable of

[1] Information adapted from http://www.fas.org/spp/military/program/com/.

transmitting and receiving communication signals with the satellites. Although a single terminal can only communicate with one satellite at a time, it normally has the capability to change from one satellite to another as required. Depending on the specific type, a terminal has the capability to support one or multiple voice and data streams. In addition, some types also have the capability to interface and control certain aspects of the satellite, such as resource allocation and antenna pointing.

The type of communication link (access control) between the space segment and the user segment is different for the uplink and the downlink. The EHF-SATCOM system uses MF-TDMA as its uplink access method and a single time-division multiplexed stream as its downlink access method. The uplink bandwidth is divided into several beams, and each beam is made up of several frequency channels. In further discussions, we assume that 32 frequency channels are available for the uplink, and each frequency channel is composed of 70 TDMA timeslots per frame.

Each terminal initiates communication (with some other terminal) by making a connection request. The connection is supported through the allocation of the commonly shared resources (i.e., set of timeslots) managed by a satellite. In an MF-TDMA satellite system, timeslots are allocated in groups, called *bursts*. Each burst is composed of a single string of contiguous timeslots over which a terminal transmits its data. A terminal transmits, to the satellite, its bursts in the assigned position of the frame according to a transmit burst time plan (BTP), and receives bursts in the assigned position of the frame, returned by the transponder, according to a receive BTP [8]. Note that a terminal may request multiple connections over time, and at any given time, a terminal may have more than one active connection.

The length of the burst (i.e., number of timeslots) depends on the modulation mode and the data rate. The EHF-SATCOM system supports seven different modulation modes and eleven different data rates for the uplink. The modulation mode determines the burst rate of the transmission, which is the rate at which symbols can be transmitted within the burst. That is, the seven modulation modes each specify a different burst rate for the terminal's uplink transmission. Note that the burst rate is in symbols per second while the data rate is in bits per second. Because the burst rate is directly affected by the BER of the connection according to the uplink budget equation, the determination of the modulation mode depends on the BER requirement (see (1) and (2)). Note that BER is one of the QoS requirements (i.e., BER and data rate) of a connection.

Given the modulation mode and the data rate, the burst length is uniquely determined using a system-specific look-up table. Although choosing a modulation mode corresponding to a higher burst rate conserves the amount of resource (i.e., number of timeslots) allocated to a connection, it also causes an increase in the BER. A higher burst rate directly translates to a higher BER (see (1) and (2)), and hence there is a trade-off between capacity and QoS in the EHF-SATCOM uplink scheme.

After the length of the burst is computed, the timeslots are allocated on the MTCS. The MTCS can be viewed as a two-dimensional array, where the rows represent frequency channels, and the columns represent timeslot indexes (see Fig. 1). When allocating timeslots on the MTCS, the following restrictions are applied:

- **Restriction 1:** The set of timeslots used by a terminal to support a given single connection must be contiguous on one frequency (i.e., must form a single burst).
- **Restriction 2:** A terminal cannot use timeslots that overlap in time to support multiple connections.

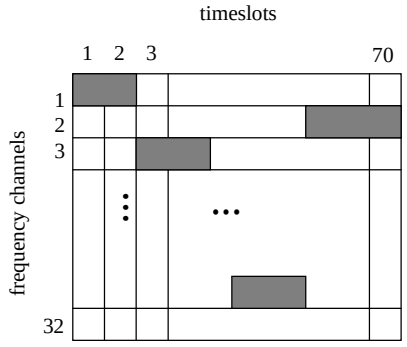

Fig. 1. MF-TDMA channel structure (MTCS). The MTCS can be viewed as a two-dimensional array, where the rows represent frequency channels, and the columns represent timeslot indexes.

These restrictions are due to the hardware and operational limitations of the EHF-SATCOM system. Earth terminals for this system employ a high power amplifier for the uplink. Nonlinearities in the amplifier create inter-modulation products when multiple carriers are present at the same time with the amplifier operating at full output power. The power in the inter-modulation products will result in reduced power in the carriers. Thus, Restriction 2 is imposed to avoid inter-modulation products [6], [7]. The reason for Restriction 1 is to ease the assignment problem for the satellite resources and simplify the routing in the payload. It follows from the two restrictions and the given channel structure that a terminal cannot be assigned more than 70 timeslots in a single frame.

When a connection is set up between two terminals via satellite, it can be established as full-duplex or half-duplex. When the full-duplex mode is used, either terminal can transmit at any time, and hence two uplink bursts must be assigned, one for each terminal. For the simulation results in Section V, we assume that the system always operates in the full-duplex mode.

B. Timeslot Calculation

To assign the appropriate number of timeslots for each connection request, we need to calculate the maximum allowable burst rate. Once the burst rate is computed, the required modulation mode can be obtained from a system-specific look-up table. Assuming that the system uses binary

DPSK, and that the required BER P_b is given, the corresponding E_b / N_0 (signal-to-noise ratio (SNR) per bit) on the uplink can be calculated as

$$P_b = \frac{1}{2} e^{-E_b / N_0}. \quad (1)$$

The SNR per bit in turn depends on various environmental and system parameters (i.e., link parameters) according to the following equation:

$$E_b / N_0 = P_t + G_t + L_f + L_r + L_c + G_c + G_r - 10\log(R_b) - 10\log(kT), \quad (2)$$

where

P_t : transmitter power in dB,

G_t : transmitter antenna gain in dB,

L_f : free-space loss in dB,

L_r : rain loss in dB,

L_c : loss due to catastrophic failure in dB,

G_c : coding gain in dB,

G_r : receiver antenna gain in dB,

R_b : burst rate in symbols per second,

k : Boltzmann's constant (1.38×10^{-23} J/K),

T : system noise temperature (assumed to be constant at 1000K).

The above equation and its counterpart for the downlink are called the *link-budget equations*.

All of the above link parameters have an impact on the behavior of the system, some greater than others. It is known that at the frequencies at which EHF-SATCOM systems operate, rain loss is the single most important parameter, aside from loss due to catastrophic failures [11]. In our model, we assume that all of these parameters, except for rain loss, are known in advance. Note that considering the rain loss as the only non-deterministic parameter does not make our model restrictive. In fact, the effect of uncertainties about the other parameters can be aggregated into the rain loss value (via (2)), and then it can be converted to an *effective rain rate* using (3) (see below). *The effective rain rate represents the aggregated effect of all the non-deterministic parameters on the SNR per bit value*.

The relation between rain loss and rain rate is given by

$$L_r = -l \times k_p \times R^{\alpha}. \quad (3)$$

Here, l is the length of the terminal to satellite path that is in rain (usually assumed to be the distance from the terminal to the freezing height along the path, if it is raining, or zero, if it is not raining). The parameter R is the rain rate described in mm/hr, and k_p and α are frequency-dependent parameters with values of 0.4 and 0.9, respectively, at 44.5 GHz (uplink frequency of the Milstar EHF-SATCOM system) [10]. The nominal values of the other parameters and the value of the rain loss, as computed above, can jointly be used to determine the burst rate (and consequently the modulation mode) required to achieve the requested BER, once the rain rate is known.

III. Resource Calculation

A. Problem Description

The resource-calculation phase deals with the problem of determining the amount of resource(s) required to provide the requested QoS. As already mentioned, we treat the timeslots as the only resource in the system. Thus, in the resource-calculation phase, we need to determine the number of timeslots required to set up a communication connection with the requested level of QoS. As explained in Section II-B, given a fixed number of timeslots, the BER and data rate depend on each other through the link-budget equations. A compromise is achieved by selecting a proper modulation mode.

We assume that the satellite system is equipped with a means of measuring the BER in the uplink and the downlink. Thus, the desired value of BER can easily be maintained as follows:

1) Observe the BER at regular intervals.
2) At every epoch, use the observed value of BER and the present burst rate to compute the burst rate required to provide the desired BER.
3) Change the modulation mode to the one that corresponds to the burst rate computed in the second step.

This scheme would be sufficient if the primary objective is to control the BER. But the connection requests require a fixed data rate as well as a guaranteed BER. If the values of the link parameters (see (2)) change during a connection's duration, this causes a corresponding change in the BER. To prevent the BER from exceeding the maximum-allowable level, while maintaining a constant data rate, the burst rate has to be constantly changed to compensate for the changes in the link parameters. With a fixed data rate, changing the burst rate requires changing the number of timeslots allocated for the connection. This means that the timeslots must be reallocated. However, timeslot reallocation in the EHF-SATCOM system is a time-consuming process, and hence this alternative is not viable.

One way to guarantee the BER and yet provide a fixed data rate is to provide some safety margin in the SNR by starting the communication in a modulation mode corresponding to a burst rate that is lower than what is required by the present BER. Thus, despite the variations in the environmental and system parameters, the safety margin should make up for the increased disturbance (i.e., any factor that is detrimental to the transmission signal), maintaining the desired BER. In the current implementation of the system, experimentally determined values are used for the parameters appearing in the link-budget equations. Specifically, as a safety margin, a 12dB allowance is added on to the E_b / N_0 computed using these parameters, and a modulation mode is selected accordingly. We will refer to this method as the *12 dB scheme*. This method is not very efficient, and one might squander a lot of timeslots, yet not always satisfy the BER requirement (and thus may

have to reconfigure the connection more often).

Here, we introduce a MMP scheme to predict the worst-case SNR per bit in terms of the effective rain rate. We then choose a modulation mode that can accommodate this predicted worst-case SNR. The principles involved in managing the uplink and the downlink are very similar. Thus, we will restrict our discussion only to the uplink. All our results are also demonstrated only for the uplink.

B. Markov Model-Based Prediction

1) Basic Approach

Given a connection request, the resource-calculation phase relies on determining the worst-case disturbance over the duration of the connection so that sufficiently many timeslots can be allocated to the connection to meet the required BER with a probability no less than some prescribed value. To determine the worst-case disturbance, we use a Markov model to characterize the disturbance process, in terms of the effective rain rate. The use of a Markov model is, in principle, not restrictive—indeed, any process of arbitrary complexity can be approximated arbitrarily well by a sufficiently large Markov model. The main caveat is the size of the model required. In the case of a noise profile that is primarily affected by weather conditions, we have found that a model with manageable size suffices. Even if the model turned out to require an unmanageable number of states, our method extends to the use of *hidden Markov models*, significantly enlarging the family of processes that can be captured with a manageable number of states. But, as pointed out above, practical considerations render such an extension of our method unnecessary. Below, we describe the specific Markov model we used to characterize the effective rain rate process, how we estimate the parameters of the model, and how we use the model to calculate the worst-case disturbance with a probability no less than some prescribed threshold value.

2) Training the Markov Model

The Markov model consists of 80 states. Each state represents the variable part of the disturbance in terms of the effective rain rate (measured in mm/hr), and whether the disturbance is increasing or decreasing. The states 0 through 39 represent the rain rates of 0 mm/hr through 39 mm/hr, and that the disturbance is either increasing or is constant. Furthermore, the states 40 through 79 represent the rain rates of 0 mm/hr through 39 mm/hr, and that the disturbance is strictly decreasing. A training profile is used to count the relative frequencies of various state transitions, which are then used to compute the transition probabilities. Thus, the training process provides us with an estimate of the probability transition matrix $\boldsymbol{P}$, where the entry P_{ij} denotes the probability of state transition from state i to state j.

3) Computing the Supremum

Assume that the duration of the connection is known in advance, and that it is N (an integer) units of time. Denote the set of states by Ω. Let $\{P_{ij}: i, j \in \Omega\}$ be the set of transition probabilities obtained from the training process. Without loss of generality, let us assume that the starting time of the connection is zero, and that the state of the system at this time is x_0. Denote the state of the system at time instant n by a random variable X_n. Thus, we have $X_0 = x_0$. Let us use the notation rainrate(x) to denote the rain rate in state x, i.e.,

$$\text{rainrate}(x) = \begin{cases} x & \text{if } 0 \le x \le 39 \\ x - 40 & \text{if } 40 \le x \le 79. \end{cases} \tag{4}$$

Given a probability threshold p^0, we wish to find the smallest value r such that

$$\Pr\{R_N \le r, R_{N-1} \le r, ..., R_0 \le r \mid X_0 = x_0\} \ge p^0, \tag{5}$$

where we use R_n to represent rainrate(X_n). If we could compute the left-hand side of (5) for any value of r (between 0 to 39), then we can easily determine the smallest r satisfying that inequality. Clearly, r has to be greater than or equal to rainrate(x_0), because otherwise the left-hand side of (5) will be zero. Thus, it suffices to consider the case where $r \ge \text{rainrate}(x_0)$. For each r, let us define a set $S(r) \equiv \{i \in \Omega : \text{rainrate}(i) \le r\}$, i.e., $S(r)$ is the set of states in which the rain rate is less than or equal to r. Then the probability on the left-hand side of (5) can easily be computed as follows:

$$\begin{aligned}
&\Pr\{R_N \le r, ..., R_0 \le r \mid X_0 = x_0\} \\
&= \Pr\{X_N \in S(r), ..., X_0 \in S(r) \mid X_0 = x_0\} \\
&= \Pr\{X_N \in S(r), ..., X_1 \in S(r) \mid X_0 = x_0\} \\
&= \sum_{x_1 \in S(r)} \Pr\{X_N \in S(r), ..., X_2 \in S(r) \mid X_1 = x_1, X_0 = x_0\} P_{x_0 x_1} \\
&= \sum_{x_1 \in S(r)} \Pr\{X_N \in S(r), ..., X_2 \in S(r) \mid X_1 = x_1\} P_{x_0 x_1} \\
&= \sum_{x_2 \in S(r)} \Pr\{X_N \in S(r), ..., X_3 \in S(r) \mid X_2 = x_2\} \sum_{x_1 \in S(r)} \left(P_{x_1 x_2} P_{x_0 x_1}\right) \\
&= \sum_{x_3 \in S(r)} \Pr\{X_N \in S(r), ..., X_4 \in S(r) \mid X_3 = x_3\} \cdot \\
&\quad \sum_{x_2 \in S(r)} \left(P_{x_2 x_3} \sum_{x_1 \in S(r)} \left(P_{x_1 x_2} P_{x_0 x_1}\right) \right) \\
&\qquad\qquad \vdots \\
&= \sum_{x_N \in S(r)} \left(\sum_{x_{N-1} \in S(r)} \left(P_{x_{N-1} x_N} \sum_{x_{N-2} \in S(r)} \left(P_{x_{N-2} x_{N-1}} \cdots \sum_{x_1 \in S(r)} \left(P_{x_1 x_2} P_{x_0 x_1}\right) \cdots \right) \right) \right).
\end{aligned}$$

Thus, the probability can be computed in $N \times |S(r)|$ computations. Given a probability threshold, we search for the smallest $r \in \{\text{rainrate}(x_0), \text{rainrate}(x_0)+1, \ldots, 39\}$ satisfying (5)—a simple binary search suffices for this purpose.

4) Computing the Number of Timeslots

Once the supremum of the effective rain rate over the connection duration is obtained, the burst rate required to satisfy the BER requirement is computed using (2) and (3). Using the computed burst rate, the corresponding modulation mode is selected. As mentioned previously, given the modulation mode and the data rate, the size of the burst (i.e., number of timeslots) is determined using a system-specific look-up table. This is the number of timeslots that will be allocated (if possible) to set up the communication connection. The method of allocating these timeslots on the MTCS, while conforming to the allocation restrictions (see Section II-A), is

described in Section IV. The process of providing guaranteed QoS connections via the EHF-SATCOM system is summarized in Fig. 2.

If the disturbance crosses the allowed safety margin (i.e., the effective rain rate of the actual disturbance becomes more than what was predicted), the actual BER exceeds the maximum-allowable BER requirement of the connection, and thus might require the connection to be reconfigured. A higher value of p^0 in (5) implies that these situations are less probable to occur. But this also costs more in terms of the number of timeslots allocated for the connection. Thus, there is a tradeoff between resource utilization efficiency and frequency of reconfiguration.

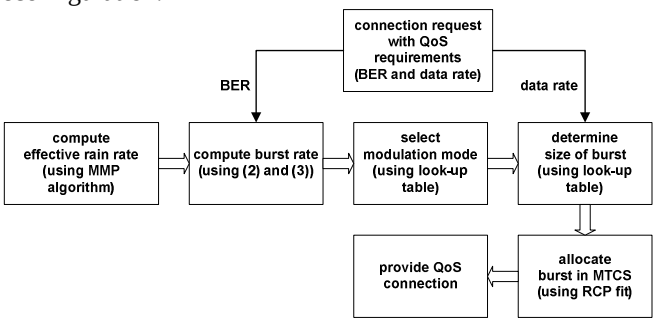

Fig. 2 Process of providing guaranteed QoS connections.

IV. Resource Allocation

A. Problem Description

After the number of timeslots has been calculated, a message is sent to the resource controller requesting these timeslots. In the resource-allocation phase, the controller invokes a resource-allocation algorithm to allocate the timeslots onto the MTCS. Allocating the timeslots efficiently is not a trivial problem—as is proven in the Appendix, it is in fact NP-complete.[2]

Of special concern is the *fragmentation* or *checkerboarding* that might prevent a burst from being allocated although the total space is sufficient for it. Due to the dynamic nature of the connection request arrivals, diversity of the burst sizes, and the allocation restrictions, frequency channels are prone to have many fragmented spaces within them. Because bursts cannot be split into smaller pieces to fit these fragmented spaces, this can result in the wastage of the uplink transmission capacity. It is apparent that reducing the fragmentation is crucial for obtaining efficiently packed channels.

The problem of allocating timeslots for the uplink can be viewed as a variant of the bin-packing problem. Most of the research efforts in this area have concentrated on acquiring close bounds on the worst-case performance of well-known packing schemes, such as first fit and best fit, applied to the static case [9]. We use first fit and best fit as benchmarks to evaluate the performance of our scheme, RCP fit. Their formal definitions are given below.

- **First fit:** Let $B_1, B_2, ...$ be the sequence of bins with each bin having a maximum capacity of C. The items $x_1, x_2, ..., x_n$ will be placed in that order starting from the first bin (i.e., B_1). To place x_i, find the least j such that B_j is filled to level $\alpha \le C - x_i$, and place x_i into B_j in the leftmost empty position (assuming that B_j's capacity is indexed from left to right). Now B_j is filled to level $\alpha + x_i$, which is less than or equal to C.
- **Best fit:** Let $B_1, B_2, ...$ be the sequence of bins with each bin having a maximum capacity of C. The items $x_1, x_2, ..., x_n$ will be placed in that order starting from the first bin. To place x_i, find j such that B_j is filled to level $\alpha \le C - x_i$, where α is as large as possible. If two or more bins with the same value of α exist, then select the bin with the smallest index. Now, place x_i into B_j in the smallest empty space large enough to fit it.

For our application, frequency channels represent the equal-capacity bins, and the bursts represent the items that need to be packed. The objective is to maximize the utilization of the MTCS, where utilization is defined as the percentage of timeslots that are actually allocated. The static model of bin packing is not directly applicable to our application, because it fails to take into account the dynamic arrivals and departures of the items. Coffman et al. formulated the *dynamic bin-packing* model and analyzed the first-fit algorithm within this context [3]. However, they did not consider the problem of managing space within a bin to reduce fragmentation. In [12], Nichols and Conklin discuss an approach specific to Milstar EHF-SATCOM systems, but their approach is limited to the static case.

B. The RCP-fit Algorithm

1) General Idea

As already mentioned, first-fit and best-fit algorithms are widely known algorithms for solving the generic bin-packing problem. These algorithms blindly pack the given items without any knowledge of the arrival statistics of the items or the special packing restrictions that might exist in a specific application. Therefore, it is possible for an algorithm to outperform these two packing schemes, if these factors are taken into consideration.

As already noted, reducing the fragmentation is crucial for obtaining efficiently packed channels. Three factors can cause fragmentation in the timeslots of the MTCS— diversity of the burst sizes, dynamic arrivals and departures of the connection requests, and the allocation restrictions. The fragmentation caused by the first two factors is unavoidable regardless of the

[2] To be precise, we prove the offline version of this problem to be NP-complete. This, however, does not prove that the online version is equally difficult to solve. The online version can be cast in the framework of Markov Decision Processes (MDP), but because of the curse of dimensionality, this approach is impractical. It is interesting though to note that some heuristic techniques—such as Hindsight Optimization—used for solving MDPs rely on finding the optimal offline solution, which is extremely difficult to find for this problem.

packing scheme. Let us consider the allocation restrictions for the timeslot-allocation problem. Restriction 1, mentioned in Section II-A, also applies to the bin-packing problem, whereas Restriction 2 is unique to our problem. Because of Restriction 2, the *p*ossible *a*llocation *s*pace for any group of bursts coming from a *s*ingle *t*erminal (PASST) is restricted to 70 timeslots, which is the length of one frame. From here on, we will denote this space simply as PASST.

Allocating timeslots according to Restriction 2 can cause fragmentation within the frequency channels, especially when the total uplink-traffic load on the system is distributed among a small number of active terminals. This case is illustrated in Fig. 3(a). In this example, there are four bursts associated with two active terminals *A* and *B*, and the bursts are allocated using first fit. The letters in the white boxes represent the terminal associated with each burst, and the shaded boxes represent the PASST for terminal *B*. We assume that the connection requests are coming from the two terminals in the order *reqA*(3), *reqB*(2), *reqA*(3), and *reqA*(1), where *reqX*(*y*) represents a request coming from terminal *X* of size *y* timeslots. In the figure, the fourth request is allocated in the fourth timeslot (i.e., fourth column) of the second frequency channel because of Restriction 2. The fragmentation caused by this allocation will prohibit any bursts longer than four timeslots from being allocated in the second channel.

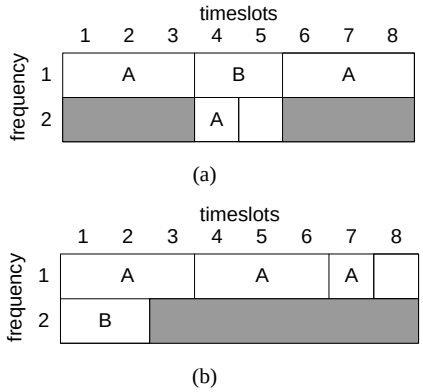

Fig. 3. Allocation example: (a) without channel reservation, and (b) with channel reservation. Four bursts associated with two active terminals are being allocated.

The fragmentation (caused by Restriction 2) described above can be avoided by grouping all the bursts that belong to the same terminal and placing them in a single frequency channel. Consider a grouping of bursts in which each frequency channel is associated with a specific terminal, that is, all bursts within a channel are from the same terminal. This grouping can be done by reserving a frequency channel for bursts associated with the same terminal, which we call *channel reservation*. This is the underlying idea behind RCP fit. An example of timeslot allocation using channel reservation is illustrated in Fig. 3(b). In the figure, the same set of bursts used in Fig. 3(a) is allocated using channel reservation. Note that in Fig. 3(b), there is no fragmentation in the second channel, allowing any burst shorter than seven timeslots to be allocated. Notice that the size of the PASST associated with terminal *B* has not changed from Fig. 3(a). However, this arrangement of bursts has allowed the PASST for terminal *B* to be contiguous, which improves the utilization of the timeslots.

We have already explained that channel reservation can be used to pack the bursts in a more space-efficient manner. However, we have implicitly assumed that $N_t \le N_c$, where N_t and N_c are the number of active terminals and the number of frequency channels, respectively. Obviously, if $N_t > N_c$, it would be impossible to reserve an exclusive channel for each of the active terminals requesting a connection. Consequently, some of the frequency channels need to be allocated with a mixture of bursts from different terminals. These *mixed channels* undermine the effectiveness of the channel reservation scheme, and should be kept to a minimum.

2) The Algorithm

Before describing the details of RCP fit, we define the following terms:

- **Channel tag:** Specifies whether a given channel is reserved, unreserved, or empty.
- **Reserved channel:** A frequency channel that is reserved for bursts coming from a specific terminal. All bursts allocated in this channel are coming from the same terminal.
- **Unreserved channel:** A frequency channel that can be shared by bursts coming from multiple terminals. This channel is characterized by a heterogeneous mix (i.e., in terms of terminals) of bursts allocated within the channel.
- **Empty channel:** A frequency channel that is completely empty. There are no bursts allocated in the channel.
- **Terminal load:** This value is used to quantify the traffic generated by each terminal. It represents the uplink-traffic load that each terminal is generating. Terminal load ρ_i is defined as

$$\rho_i = \frac{d_i \cdot l_i}{\tau_i}, \tag{6}$$

where

d_i: mean duration of the connections from terminal i (in frames);

l_i: mean burst length per frame of the connections from terminal i (in timeslots/frame);

τ_i: mean interarrival time of the connection requests coming from terminal i (in frames).

In (6), d_i and τ_i are measured in units of frames, and l_i is measured in units of timeslots per frame. Hence, terminal load is a quantity measured in (timeslots/frame). Because a frame is of a fixed duration, this measure is equivalent to (timeslots/time).

If $N_t > N_c$, some bursts must be allocated in an unreserved channel that is already occupied by bursts coming from different terminals. The unreserved channels undermine the effectiveness of channel reservation, and contribute to the fragmentation of the MTCS. An unreserved channel is created

only if the following conditions are satisfied when trying to allocate a burst (see Fig. 4):

- There is no reserved channel that is associated with the terminal of the burst.
- There is no empty channel.
- There is no (previously created) unreserved channel that has enough space to accommodate the burst.

When an unreserved channel must be created (to allocate the burst), it is created by selecting a reserved channel and changing it into an unreserved channel. To minimize the number of unreserved channels, the criterion for selecting the reserved channel (which will be changed into an unreserved channel) is based on the traffic load created by the terminal associated with each reserved channel. We assume that the traffic load is unequally distributed among terminals, and that the system can detect these differences.[3] It is likely that terminals sharing the same uplink resource (i.e., terminals within the same beam or terminals in different beams that are using the same satellite uplink) will each generate different amounts of uplink traffic. For example, certain terminals might be sending high-resolution images, requiring large amounts of channel resources, while other terminals might be sending text messages that require much less resources. To quantify the traffic generated by each terminal, a quantity called *terminal load* (see (6)) is calculated for each terminal. After the terminal load value is calculated for each terminal associated with a reserved channel, a reserved channel is selected whose terminal has the smallest terminal load. This reserved channel is changed to an unreserved channel (by changing the channel tag), and the burst is allocated within it.

In Fig. 4, we describe the steps of RCP fit using a flowchart. Note that in the timeslot allocation step, once a frequency channel is selected, the burst is allocated in the smallest empty space available within the frequency channel that is large enough to fit it. To illustrate how the RCP-fit algorithm can actually be used to allocate bursts, here we give an example allocation scenario. We assume the following:

- Channel structure is MF-TDMA (16 timeslots × 4 channels).
- Connection requests are coming from five terminals—*A*, *B*, *C*, *D*, and *E*.
- The order of the connection requests is *reqA*(3), *reqB*(8), *reqA*(8), *reqC*(2), *reqD*(6), *reqE*(2), *reqA*(5), *reqC*(4), and *reqE*(8).
- The order of the terminal load generated by each terminal (from largest to smallest) is *A*, *B*, *C*, *D*, and *E*.
- For simplicity, we consider the static case only (i.e., connection terminations are not considered).

The first request is accommodated by placing a burst of size three in timeslots one through three of Channel 1 (see Fig. 5(b)). The first channel's tag is changed to "reserved," which means that this channel should be packed only with terminal *A*'s bursts. The second burst is placed in Channel 2, and this channel is reserved for terminal *B*. Similar procedures are followed for the fourth and fifth bursts. The third burst is placed in the first channel without any change in the channel tag because this channel has already been reserved for terminal *A*. To accommodate the sixth request coming from terminal *E*, one of the reserved channels must be changed to an "unreserved" channel because all the channels have already been reserved for other terminals. According to the RCP-fit algorithm, we choose Channel 4, which is reserved for terminal *D*, and allocate terminal *E*'s burst in this channel. Recall that terminal *D*'s terminal load value is smaller than that of terminal *A*, *B*, or *C*. The last burst (for terminal *E*) is allocated in Channel 4 because this channel is unreserved and has enough space to accommodate the burst.

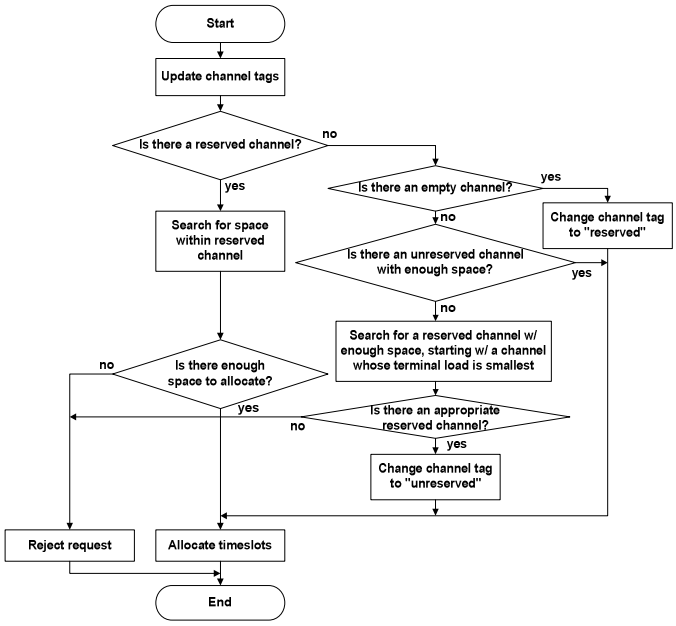

Fig. 4 Flowchart for RCP fit.

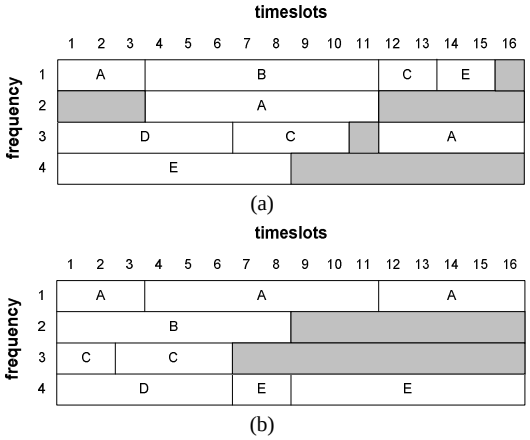

Fig. 5. Allocation example: (a) first-fit, (b) RCP-fit.

[3] When a terminal has traffic to send, it will request a channel access via the resource controller. The resource controller can keep a historical record of the connection requests attributed to each terminal, including parameters such as connection duration and burst length of the connection. Using this record, the resource controller can keep track of the resources allocated to each terminal, and estimate the traffic load generated by each terminal.

Fig. 5(a) and 5(b) show the MTCS after the bursts have been allocated using the first-fit and RCP-fit algorithms, respectively. The shaded regions represent empty timeslots. The figures clearly illustrate that packing with RCP fit causes less fragmentation. We claim that packing bursts via RCP fit results in improved utilization of the MTCS by reducing the fragmentation (compared to first fit and best fit). Simulation results of Section V support this claim.

V. Simulation Results

A. Resource Calculation Using Markov Model-Based Prediction

1) Simulation Details and Performance Measures

The simulations were performed on computer-generated Markov and non-Markov disturbance profiles. We performed several different simulations using Markov as well as non-Markov profiles. Each simulation used two different profiles (with the same probability distribution)—one for the training phase and one for the prediction phase. Profiles used in different simulations had different distributions signifying different "coarseness levels" in the disturbance profiles. We say that a profile is *coarser* than the other if the state-transition probabilities for the former profile are higher than those for the latter. The training phase used profiles that were 1,000,000 sample points (spaced at two seconds) in length. In the prediction phase, 100,000 connection requests were generated according to a Poisson-arrival process with a mean interarrival time of 20 seconds. The connection-holding times were distributed uniformly between zero and 400 seconds (these were approximated to the closest number of sample points). The simulations were used to compare the performance of the MMP scheme with that of the 12 dB scheme currently implemented in the Milstar EHF-SATCOM system. For comparing the performance of the two techniques, the following two measures were used.

- **Slot allocation factor:** Let $S_i^{\min}$ be the minimum number of timeslots required to satisfy the BER requirement of connection *i* without reconfiguring the connection. Note that this can be determined by observing the disturbance profile, but only after the connection has been completed. Let S_i^A be the number of timeslots allocated by algorithm *A* for connection *i* (where *A* represents either the 12 dB scheme or the MMP scheme). Then the slot allocation factor (SAF) of algorithm *A* is defined as

$$SAF(A) = \frac{\sum_{i \in \{\text{all calls}\}} \left| S_i^A - S_i^{\min} \right|}{\sum_{i \in \{\text{all calls}\}} S_i^{\min}}.$$

  Intuitively, *SAF*(*A*) indicates the normalized number of timeslots wasted by algorithm *A* on the average. An algorithm with a lower SAF value wastes fewer number of timeslots. As resource-allocation efficiency is of utmost concern over wireless links, we believe that this performance metric mirrors the resource constraints faced by most satellite systems.

- **Fraction of instants the QoS is not satisfied:** The performance metric SAF defined above provides a measure of algorithm efficiency. An algorithm with low SAF is more efficient compared to an algorithm with larger SAF, as it wastes fewer timeslots. But, SAF gives only a one-sided view of the performance of the algorithm. If one compares two algorithms purely based on SAF, then an algorithm allocating *no timeslots* would be the best. But clearly this is not an acceptable solution. What we are interested in is an algorithm that attempts to meet the QoS requirements of the users as much as possible, and yet is frugal with the timeslots. Thus, we define our second metric—fraction of instants the QoS is not satisfied. This metric is defined as the number of connections for which the BER requirement is not satisfied, divided by the total number of connections. This criterion measures the effectiveness of the algorithm in terms of providing the required BER level.

2) Simulation Results and Discussion

For brevity, we provide simulation results only for two disturbance profiles. Figs. 6 and 7 show the performance of the two schemes for the case of a moderate, non-Markov profile, whereas Figs. 8 and 9 show the results for a very coarse, non-Markov profile.

We can see that for Fig. 6, our scheme outperforms (i.e., results in lower SAF values compared to) the 12 dB scheme for probability thresholds lower than 0.8. For Fig. 7, our scheme always outperforms (i.e., results in lower values for the fraction of instants the QoS is not satisfied compared to) the 12 dB scheme. On the other hand, for Fig. 8, the 12 dB scheme always outperforms our scheme in terms of SAF. But it should be noted that the fraction of instants QoS is not satisfied is as high as 0.26 for the 12 dB scheme using the same profile (see Fig. 9). This value is intolerably high, as connection reconfiguration can take as long as 40 seconds in a Milstar EHF-SATCOM system. Thus, the MMP scheme achieves a compromise between bandwidth efficiency and frequency of reconfiguration, as opposed to the 12 dB scheme, which does not achieve this compromise. Also, the MMP scheme has the advantage of being able to tune its parameter (i.e., probability threshold) to adjust to the disturbance profiles.

B. Resource Allocation Using RCP Fit

The performance of RCP fit was simulated using a system modeled after the Milstar EHF-SATCOM system described in Section II. The following assumptions were made.

- Connection requests arrive according to a Poisson process.
- Duration of a connection is exponentially distributed.
- Minimum required BER for each connection is fixed at 10^{-5}.
- Data rate for each connection is randomly picked from the eleven rates supported by the EHF-SATCOM system, with equal probability.
- Parameters (i.e., transmitter power, transmit/receive antenna gain, free space loss, loss due to catastrophic failure, coding gain, and system noise temperature) of the uplink budget equation (see (2)) are fixed at system-specific nominal values.

- Rain loss is calculated using (3), where the rain rate values are obtained from a simulated rain profile.
- Channel structure: MF-TDMA (70 timeslots × 32 channels)
- The system operates only in full-duplex mode.

Figs. 10 and 11 compare the three packing algorithms (i.e., first fit, best fit, and RCP fit) when the connection requests are *unbiased*. Here, unbiased means that all terminals generate the same amount of uplink traffic. For all three packing algorithms, the algorithm is applied without violating the allocation restrictions of Section II-A. For comparing the performance of the three packing schemes, the following two measures were used.

- **Utilization:** Utilization is defined as the percentage of timeslots that are actually being allocated.
- **Timeslot rejection ratio (TRR):** The timeslot rejection ratio is defined as

$$\frac{100}{T}\int_0^T \frac{\min(N_r^t, N_{\max}) - N_a^t}{\min(N_r^t, N_{\max})} dt,$$

where N_r^t is the number of requested timeslots at time t, N_a^t is the number of allocated timeslots at time t, $N_{\max}$ is the maximum capacity of the MTCS in terms of timeslots, and T is the observation interval. Intuitively, TRR indicates the normalized number of timeslots rejected because of the fragmentation in the MTCS.

Figs. 10 and 11 show the utilization and the TRR versus load factor with the number of terminals requesting a connection on the uplink fixed at thirty. The load factor is the mean duration of the connections divided by the interarrival time of the connection requests. It can be seen from the curves that RCP fit outperforms the other packing algorithms in terms of the performance measures mentioned above. When the load factor is 100, packing the timeslots with RCP fit instead of best fit increases the utilization by 17% and decreases the TRR by 3%. Note that all three packing schemes result in relatively low utilization and high TRR. These results are partly caused by the fact that the data rate-BER combination of some of the connections requires a long burst length, some as long as 64 timeslots in length, which is very difficult to allocate due to any existing fragmentation in the MTCS.

Figs. 12 and 13 compare the three packing algorithms when the connection requests are *biased*. Here, biased means that certain terminals have greater terminal load values than those of others, and the *bias factor* is used to quantify this value. For example, if the bias factor of a terminal is 24, then it means that the terminal has a terminal load value that is 24 times greater than that of unbiased terminals.

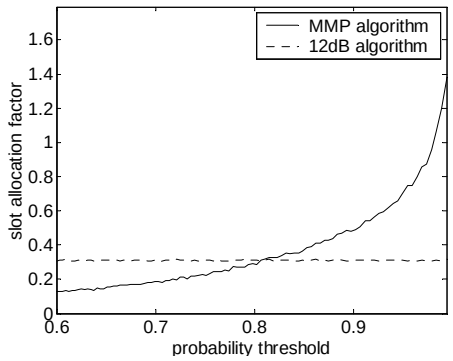

Fig. 6. SAF in the case of a moderate non-Markov profile.

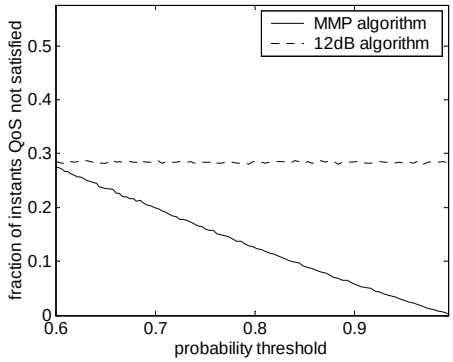

Fig. 7. Fraction of instants QoS is not satisfied for the case of a moderate non-Markov profile.

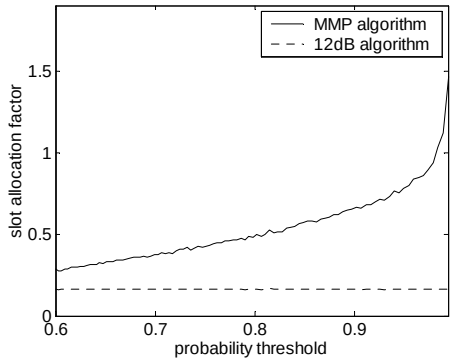

Fig. 8. SAF in the case of a very coarse non-Markov profile.

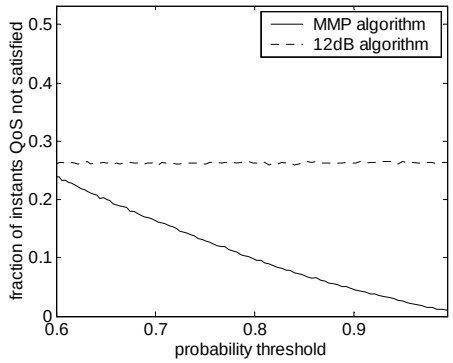

Fig. 9. Fraction of instants QoS is not satisfied for the case of a very coarse non-Markov profile.

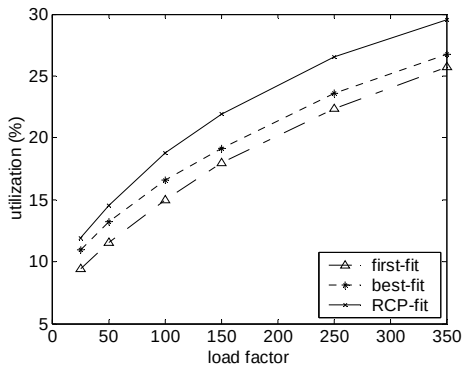

Fig. 10. Utilization for unbiased requests, 30 terminals.

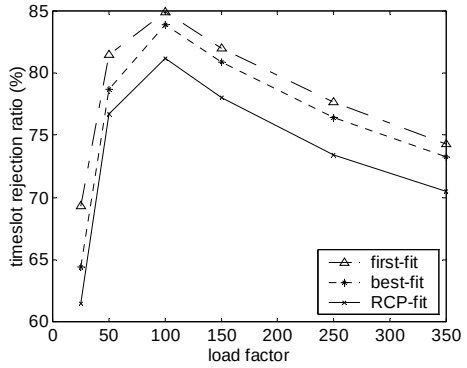

Fig. 11. Timeslot rejection ratio for unbiased requests, 30 terminals.

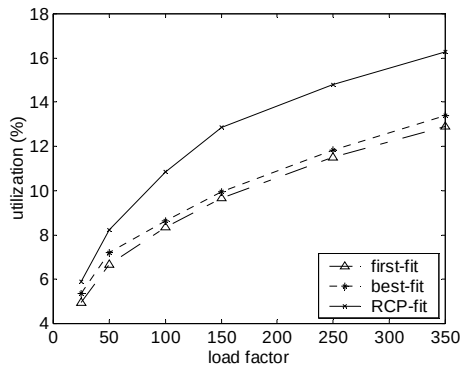

Fig. 12. Utilization for biased requests, 30 terminals, two biased terminals, both with bias factor = 24.

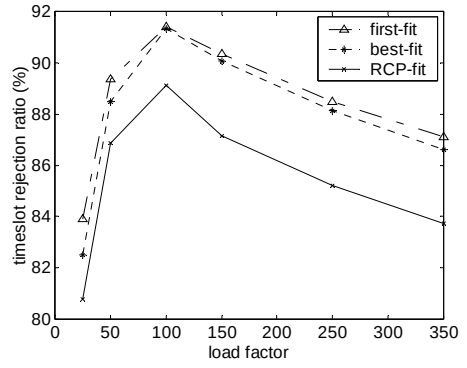

Fig. 13. Timeslot rejection ratio, 30 terminals, two biased terminals, both with bias factor = 24.

Figs. 12 and 13 show the plots for utilization and TRR versus load factor with 30 terminals. Here, the number of biased terminals is fixed at two, and the bias factor for these terminals is set to 24. We can see that the relative performance improvement obtained by RCP fit is increased when the connection requests are biased (compared with Figs. 10 and 11). For example, a utilization improvement of 29% is obtained when RCP fit is used instead of best fit at a load factor of 100. This implies that RCP fit is especially effective when a few terminals heavily dominate the uplink-traffic load. Comparing Figs. 10 and 11 with Figs. 12 and 13, we can see that irrespective of the packing scheme, the overall packing efficiency is decreased when the connection requests are biased.

VI. Conclusions

We described a scheme for providing QoS connections over MF-TDMA satellite systems. We divided the problem into two parts—resource calculation and resource allocation. For the resource calculation part, we used a Markov model-based prediction scheme and compared its performance with the scheme currently implemented in the Milstar EHF-SATCOM systems. For comparing the performance of these schemes, we used two performance measures. We demonstrated that for a moderate disturbance profile, there exists a range of probability thresholds for which our scheme performs better than the currently implemented scheme in terms of both performance measures. Moreover, for a very coarse profile, we showed that our scheme attains a compromise between frequency of reconfiguration and resource-utilization efficiency.

For the resource allocation part, we described a novel packing algorithm that can be used to allocate timeslots in the uplink of an EHF-SATCOM system. The packing efficiency of the proposed algorithm was benchmarked using simulation results; we compared the utilization and the TRR with two other packing schemes—best fit and first fit. The simulation results showed that RCP fit performs better than the other two packing schemes in both the cases considered (biased and unbiased connection requests). Furthermore, the proposed algorithm is especially effective when the uplink-traffic load is heavily dominated by a small number of terminals.

Our results were obtained using specifications and parameters of an actual Milstar EHF- SATCOM system. The proposed algorithms are applicable to conventional satellite systems employing the MF-TDMA uplink access method with similar specifications.

Appendix: NP-Completeness of the Dynamic Resource-Allocation Problem

We prove that the dynamic version of the resource-allocation problem (DRAP) is NP-complete. We will do so by reducing the bin-packing problem (BPP) to the static version of the resource-allocation problem (SRAP) and then by reducing the SRAP to DRAP. To proceed with the proof, these problems need to be defined formally. As is commonly done in

NP-completeness theory, we will cast these problems as decision problems (problems with yes/no answers) [5]. To define these problems formally, we start with some basic concepts used in the construction of the proof.

In Section II, we described the MF-TDMA channel structure (MTCS), and defined the notion of a burst. Bursts need to be allocated in a string of contiguous, empty timeslots on a single frequency channel. Such a string is denoted by a three-tuple (*f*, *p*, *l*), where *f* is the frequency channel, *p* is the position in the frequency channel, and *l* is the length of the string. Thus, a collection *E* of strings of contiguous timeslots is a set of such three-tuples.

A burst is completely characterized by its size and terminal, and thus is represented by the ordered pair (*s*, *t*), where *s* is an integer representing the burst size, and *t* is the terminal number. Thus, a collection *B* of bursts is a set of such ordered pairs.

Given a collection *B* of bursts, and a collection *E* of empty, contiguous timeslots, an allocation of bursts specifies the timeslots on the MTCS corresponding to *E*, to which each of the bursts from *B* are mapped. The allocation is valid if the following three conditions are satisfied:

- Assigns all the bursts in the timeslots corresponding to the strings in *E*.
- Does not assign the same timeslot to more than one burst.
- Satisfies Restrictions 1 and 2 described in Section II.

Now, we extend the concept of a valid allocation to the dynamic model (i.e., allocation with time considerations). A timed burst is a three-tuple (*s*, *t*, *d*), where *s* is an integer representing the burst size, *t* is the terminal number, and *d* is the duration. Note that the unit of *s* is in terms of timeslots, and the unit of *d* is in terms of frames. Thus, if the burst arrives at time *i*, then it is active till time $i + d$. A collection B_i of such three-tuples denotes a set of bursts that arrive at time *i*. The allocation of the bursts in B_i onto the MTCS is represented by A_i. Given a collection *E* of empty, contiguous timeslots, a sequence $\{B_i\}$ of collections of timed bursts, and a sequence of allocations $\{A_i\}$, we say that the allocation sequence is valid if the following three conditions are satisfied:

- All the bursts from the sequence $\{B_i\}$ are allocated in the timeslots corresponding to the strings in *E*.
- Any timeslot is allocated to at most one active burst at any given time.
- At any given time, all the active bursts satisfy Restrictions 1 and 2 described in Section II.

Using the basic concepts described above, we define the various problems as follows:

Definition 1: *The Bin-Packing Problem* (*BPP*)
Given *n* equal-capacity bins with integer capacity, and a set of integer-sized items, is it possible to fit all the items in the *n* bins?

Definition 2: *The Static Resource-Allocation Problem* (*SRAP*)
Given a collection of contiguous, unoccupied timeslots, say *E*, in the MTCS, and a set of bursts, say *B*, does there exist a valid allocation?

Definition 3: *The Dynamic Resource-Allocation Problem* (*DRAP*)
Given a collection of contiguous, unoccupied timeslots, say *E*, and a sequence of sets of timed bursts, say $\{B_i\}$, does there exist a valid sequence of allocations?

Using these definitions, we will prove that DRAP is NP-complete by first reducing the BPP to SRAP, and then by reducing SRAP to DRAP. The proofs follow the procedures for proving NP-complete problems as outlined in [4].

***Lemma 1*:** *SRAP is NP-complete*.

***Proof*:** It is easy to see that SRAP$\in$ NP. Assume that we are given a set of *k* frequency channels $\{ F_i, i = 1, ..., k \}$ with a set of bursts $B = \{ (s_i, t_i) : i = 1, \ldots, m \}$, and an allocation. We can verify that the allocation is valid, and this verification can be performed in a straightforward manner in polynomial time.

Next we show that SRAP is NP-hard by showing that BPP, which is NP-complete [9], is polynomial-time reducible to SRAP. As the first step, we take an instance of BPP, where there are *k* empty bins $\{N_i, i = 1, \ldots, k\}$ with equal capacity *L*. We transform this instance of BPP into an instance of SRAP with *k* frequency channels having *kL* timeslots each. Fig. 14 shows such an instance of SRAP. In this instance of SRAP, frequency channel *i* has the following structure: timeslots $(i-1) \times L + 1$ through $i \times L$ are unoccupied; all the other timeslots on this frequency channel are occupied by other bursts. Thus, the shaded area in Fig. 14 corresponds to the timeslots that have already been assigned to other bursts, and the unshaded area along the diagonal corresponds to the unoccupied timeslots on each of the frequency channels. The unoccupied timeslots are deliberately positioned along the diagonal of the MTCS so that the allocation is free of the restriction that a terminal cannot use timeslots that overlap in time to support multiple connections. To complete the transformation, we need to specify the burst sizes that need to be allocated on these timeslots and the terminal numbers for each of the bursts. We set these burst sizes equal to the item sizes from the BPP, and assign an arbitrary terminal number to each of the bursts. Assignment of terminal numbers to bursts can be arbitrary because of the diagonal configuration of the unoccupied timeslots in the frequency channels (see Fig. 14). Clearly, this transformation can be performed in polynomial time.

Now all that needs to be shown is that a solution to the above-mentioned instance of SRAP exists if and only if a solution to the original instance of BPP exists, and that one can be computed from the other in polynomial time. In fact, from the way we constructed the instance of SRAP, there exists a natural bijection between the set of items in the BPP and the set of bursts in the SRAP. Also, as the number of bins is the same as the number of frequency channels, we can define a bijective mapping between these two sets as well. Using these bijections, one can compute a solution to an instance of BPP from the corresponding solution of SRAP and vice versa. Also, it can be readily seen that one solution is valid if and only if

the other is also valid.

■

Now, given that SRAP is NP-complete, we can show that the dynamic resource-allocation problem is also NP-complete.

Theorem 1**:** *DRAP is NP-complete*.

Proof**:** As before, it is easy to see that DRAP∈ NP. Given a collection of empty, contiguous timeslots, *E*, a sequence of timed bursts, $\{B_i\}$, and a sequence of allocations, $\{A_i\}$, one can easily verify whether the allocation sequence is a valid allocation sequence in a straightforward manner in polynomial time.

Because SRAP was shown to be NP-complete from Lemma 1, all that remains to be shown is that every instance of SRAP (i.e., a set of empty timeslots and a set of bursts) can be converted to an equivalent instance of DRAP (a set of empty timeslots and a sequence of sets of timed bursts). Let (E,B) be any instance of SRAP, where *E* is the set of empty timeslots and *B* is the set of bursts. Corresponding to this instance of SRAP, we create the following instance of DRAP, which is denoted by $(E,\{B_i\})$. The set of empty timeslots *E* is same as that in the instance of SRAP. From the set of untimed bursts *B*, we create a set of timed bursts B_1 in such a way that every burst in B_1 has duration of one. Thus, for every burst $(s,t)\in B$, we add a timed burst $(s,t,1)$ to B_1. Then, our instance of DRAP is given by $(E,\{B_1,\phi,\phi,\ldots\})$, where ϕ represents the empty set. It can easily be seen that solution to this instance of SRAP exists if and only if solution to the instance of DRAP exists, and that one can be computed from the other in polynomial time. Thus, DRAP is NP-complete.

■

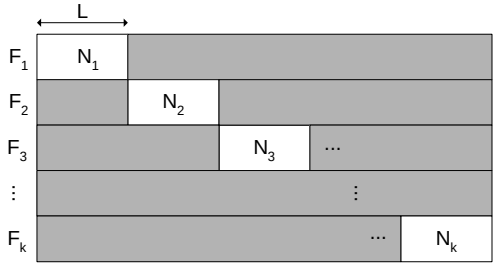

Fig. 14. An instance of BPP mapped to an instance of SRAP.

ACKNOWLEDGMENT

The authors are grateful to the Editor and to the anonymous reviewers whose comments helped to improve this manuscript. The authors would also like to thank Ruiliang Chen for his helpful suggestions.

REFERENCES

[1] G. Açar and C. Rosenberg, "Algorithms to compute bandwidth on demand requests in a satellite access unit," in *Proc. Fifth Ka-Band Utilization Conference*, Oct. 1999.

[2] E. G. Coffman Jr., K. So, M. Hofri, and A. C. Yao, "A stochastic model of bin-packing," *Information and Control*, vol. 44, no. 2, 1980, pp. 105–115.

[3] E. G. Coffman Jr., M. R. Garey, and D. S. Johnson, "Dynamic bin packing," *SIAM Journal on Computing*, vol. 12, no. 2, May 1983, pp. 227–258.

[4] T. H. Corman, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, McGraw-Hill, New York, NY, 1990.

[5] M. R. Garey and D. S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, New York, NY, 1979.

[6] R. D. Gaudenzi, "Payload non-linearity impact on the Globalstar forward link multiplex. Part I: Physical layer analysis," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 3, May 1999, pp. 960–976.

[7] D. M. Goebel, R. R. Liou, W. L. Menninger, X. Zhai, and E. A. Adler, "Development of linear traveling wave tube amplifiers for telecommunications applications," *IEEE Transactions on Electron Devices*, vol. 48, no. 1, Jan. 2001, pp. 74–81.

[8] T. T. Ha, *Digital Satellite Communications*, 2nd ed., McGraw-Hill, Singapore, 1990.

[9] D. S. Johnson, A. Demers, J. D. Ullman, M. R. Garey, and R. L. Grahm, "Worst-case performance bounds for simple one-dimensional packing algorithms," *SIAM Journal on Computing*, vol. 3, no. 4, Dec. 1974, pp. 299–325.

[10] S. D. Jones, Johns Hopkins University Applied Physics Laboratory, *Internal Technical Memorandum: EHF Network QoS Derivations*, Mar. 1999, 6 pp.

[11] A. Mathur, T. M. Nguyen, and G. Goo, "Propagation effects on the wideband gapfiller communication link," in *Proc. IEEE Military Communications Conference (MILCOM), Section U19.4*, Oct. 2000.

[12] R. A. Nichols and R. E. Conklin Jr., "Uplink packing of army Milstar services," in *Proc. IEEE Military Communications Conference (MILCOM)*, Oct. 1998.

[13] J.-M. Park, U. Savagaonkar, E. K. P. Chong, H. J. Siegel, and S. D. Jones, "Efficient resource allocation for QoS channels in MF-TDMA satellite systems," in *Proc. IEEE Military Communications Conference (MILCOM)*, Oct. 2000, pp. 645–649.

Jung-Min Park (M'03) was born in Seoul, South Korea, in March 1972. He received the B.S. and M.S. degrees both in electronic engineering from Yonsei University, Seoul, South Korea, in 1995 and 1997, respectively; and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2003.

He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA. From 1997 to 1998, he worked as a cellular systems engineer at Motorola Korea Inc. His current interests are in network security, applied cryptography, and networking. More details about his research interests and publications can be found at http://www.ecpe.vt.edu/faculty/park.html

Dr. Park is a member of the Association for Computing Machinery (ACM) and the Korean-American Scientists and Engineers Association (KSEA). He was a recipient of a 1998 AT&T Leadership Award.

Uday Savagaonkar received the Master of Technology degree in Electrical Engineering from Indian Institute of Technology, Mumbai, India in July 1998 and PhD degree in Electrical Engineering from Purdue University, West Lafayette, IN, USA in December 2002.

He joined Intel Corporation as a Senior Product Development Engineer in October 2002, and currently he works as a Senior Network Software Engineer at the Intel Communications Technologies Lab in Hillsboro, OR, USA. His research interests include network security, self-healing networks, and performance analysis of secure architectures.

Edwin K. P. Chong (S'86, M'91, SM'96, F'04) received the B.E.(Hons.) degree with First Class Honors from the University of Adelaide, South Australia, in 1987; and the M.A. and Ph.D. degrees in 1989 and 1991, respectively, both from Princeton University, where he held an IBM Fellowship.

He joined the School of Electrical and Computer Engineering at Purdue University in 1991, where he was named a University Faculty Scholar in 1999, and was promoted to Professor in 2001. Since August 2001, he has been a Professor of Electrical and Computer Engineering and a Professor of

Mathematics at Colorado State University. His current interests are in communication networks and optimization methods. He coauthored the recent best-selling book, An Introduction to Optimization, 2nd Edition, Wiley-Interscience, 2001.

Prof. Chong was on the editorial board of the IEEE Transactions on Automatic Control, and is currently an editor for Computer Networks. He is a Fellow of the IEEE, and served as an IEEE Control Systems Society Distinguished Lecturer. He received the NSF CAREER Award in 1995 and the ASEE Frederick Emmons Terman Award in 1998. He was a co-recipient of the 2004 Best Paper Award for a paper in the journal Computer Networks.

Howard Jay Siegel (M'77, SM'82, F'90) received a B.S. degree in electrical engineering and a B.S. degree in management (1972) from the Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, and the M.A. (1974), M.S.E. (1974), and Ph.D. degrees (1977) from the Department of Electrical Engineering and Computer Science at Princeton University, Princeton, New Jersey.

In August 2001, he was appointed the George T. Abell Endowed Chair Distinguished Professor of Electrical and Computer Engineering at Colorado State University (CSU), Fort Collins, Colorado, where he is also a Professor of Computer Science. In December 2002, he became the first Director of the university-wide CSU Information Science and Technology Center (ISTeC). From 1976 to 2001, he was a professor in the School of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana. He has co-authored over 300 published technical papers on parallel and distributed computing and communications, has edited/co-edited eight volumes, and wrote the book Interconnection Networks for Large-Scale Parallel Processing (second edition 1990). His research interests include heterogeneous parallel and distributed computing, communication networks, parallel algorithms, parallel machine interconnection networks, and reconfigurable parallel computer systems. He has been an international keynote speaker and tutorial lecturer, and has consulted for industry and government. More information is available at http://www.engr.colostate.edu/~hj.

Prof Siegel is an ACM Fellow, was a Coeditor-in-Chief of the Journal of Parallel and Distributed Computing, and was on the Editorial Boards of both the IEEE Transactions on Parallel and Distributed Systems and the IEEE Transactions on Computers. He was Program Chair/Co-Chair of three major international conferences, General Chair/Co-Chair of six international conferences, and Chair/Co-Chair of five workshops. He has served as a member of over 40 conference and workshop program committees. He has served as Chair of the IEEE Computer Society Technical Committee on Computer Architecture (TCCA) and Chair of the ACM Special Interest Group on Computer Architecture (SIGARCH). He was an "IEEE Computer Society Distinguished Visitor" and an "ACM Distinguished Lecturer," giving invited seminars about his research around the country. He is a member of the Eta Kappa Nu electrical engineering honor society, the Sigma Xi science honor society, and the Upsilon Pi Epsilon computing sciences honor society.

Steven D. Jones (M'85) received the B.S.E.E. degree from the University of Maryland, College Park, in 1981, and the M.S.E.E. degree from The Johns Hopkins University in 1985. He joined the Johns Hopkins Applied Physics Lab in 1981, where he currently holds the position of Principal Professional Staff and is the lead engineer for several military communications programs with the Communication Systems and Network Engineering Group in the Power Projection Systems Department. His experience has involved the architecture, development, analysis, simulation, and testing of satellite and terrestrial communications systems.