Radio Resource Management Optimization of Flexible Satellite Payloads for DVB-S2 Systems

Giuseppe Cocco, *IEEE, Member*, Tomaso De Cola, *IEEE, Member*, Martina Angelone, Zoltan Katona, and Stefan Erl

Abstract

The increasing demand for high-rate broadcast and multicast services over satellite networks has pushed for the development of High Throughput Satellites characterized by a large number of beams (e.g., more than 100). This, together with the variable distribution of data traffic request across beams and over time, has called for the design of a new generation of satellite payloads, able to flexibly allocate bandwidth and power. In this context, this paper studies the problem of radio resource allocation in the forward link of multibeam satellite networks adopting the Digital Video Broadcasting - Satellite - Second Generation (DVB-S2) standard. We propose a novel objective function with the aim to meet as close as possible the requested traffic across the beams while taking fairness into account. The resulting non-convex optimization problem is solved using a modified version of the simulated annealing algorithm, for which a detailed complexity analysis is presented. Simulation results obtained under realistic conditions confirm the effectiveness of the proposed approach and shed some light on possible payload design implications.

I. Introduction

The advent of High Throughput Satellite (HTS) systems [1], [2] has revolutionized the concept of satellite communications in that new satellite systems operating in the Ka frequency band (and above) are being designed so as to provide geographical coverage through a large number of beams. Such dramatic change has started upon the ever-increasing user demand for broadcast/multicast services characterized by high data rates and reliability performance. In order to meet these requirements, a natural technology candidate is the Digital Video Broadcasting - Satellite - Second Generation (DVB-S2) standard [3], which is nowadays one of the most widespread and preferred options by satellite broadcasters in the forward link[1].

In spite of the attractive performance figures that can be attained by DVB-S2 (e.g., in terms of spectral efficiency), the problem of optimally allocating bandwidth to beams and optimally operating the payload from a power perspective according to the amount of requested traffic is still not completely solved. This is because of the large number of variables that play an important role in the resulting radio resource allocation problem. Such problem has been often addressed from a ground segment viewpoint, by proposing optimization frameworks able to take into account propagation impairments (e.g., rain) and

G. Cocco is with the LTS4 Signal Processing Laboratory and the Laboratory of Intelligent Systems, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland, (e-mail: giuseppe.cocco@epfl.ch). G. Cocco was previously with the Institute of Communications and Navigation, German Aerospace Center (DLR), Oberpfaffenhofen 82234, Germany

T. De Cola, Z. Katona and S. Erl are with the Institute of Communications and Navigation, German Aerospace Center (DLR), Oberpfaffenhofen 82234, Germany (e-mail: {tomaso.decola, zoltan.katona, stefan.erl}@dlr.de)

M. Angelone is with the Communications and TT&C Systems and Techniques Section, European Space Agency, ESTEC, Noordwijk 2201, The Netherlands (e-mail: martina.angelone@esa.int)

[1]The standard has been recently extended in the DVB-S2X [4], which includes new features and yields improved performance with respect to the previous version.

interference contribution from other beams (e.g., co-channel interference (CCI)). For instance, reference [5] addresses the problem from a scheduling viewpoint, allocating different modulation and coding schemes (ModCods) to the different satellite beams. However, the complex characteristics of data traffic (time- and space-correlation, heavily depending on the specific geographic area) have always represented a formidable obstacle against the derivation of closed-form solutions, hence requiring the introduction of approximated models or the use of numerical optimization techniques.

On the other hand, the recent years have also witnessed an evolution of satellite system concepts from a space segment viewpoint, which also have an important impact on the resource allocation problem [6]. Specifically, more sophisticated payload designs have been introduced [7], so as to cope with the time and geographic variations of the traffic requested by each beam. This mainly resulted in two possible design options, namely beam-hopping and flexible payloads. The former makes use of a time-slotted illumination window so that a sequence of beam illumination and the illumination time assigned to each beam can be designed according to the traffic demands and the antenna radiation pattern. The latter option makes use of a dual approach, consisting in allocating bandwidth or power to the beams in relation to the requested traffic. The effectiveness of both options heavily depend on the specific payload design (e.g., number of traveling-wave tube amplifiers (TWTA) and structure of the payload connection matrix) and on the constraints (e.g., mass and available power) imposed by state-of-the-art technology. In [8], [9] the problem of time/beam allocation is studied in the presence of traffic asymmetry. In the papers a closed form solution for the optimal resource allocation in a simplified setup with no interference is derived for two different utility functions, aiming at matching the requested bitrate and maximizing the product of the ratios between the offered and requested capacity across the beams, respectively.

In [9], [10] the advantages of multi-beam with respect to single beam satellite systems are studied considering different performance metrics. Specifically, the optimal power allocation is derived for two different objective functions, one leading to throughput maximization and the other related to fairness. Although these two papers offer interesting hints on the problem of resource allocation, the validity of the results is limited by the assumption of no co-channel interference, which is instead removed in [11]. In the paper a phased array antenna is assumed at the satellite and call-admission control schemes are investigated. Differently from the approaches adopted in the papers mentioned above, the studies presented in [12], [13] explore the benefits of power allocation. In particular, a two-stages sub-optimal algorithm is applied to solve a non-convex optimization problem, the solution of which gives some insights about the relations between power allocation and offered traffic on the forward link of satellite networks. Finally, beam-hopping and flexible systems are compared in [14], where the latter implements a non-uniform bandwidth allocation and makes use of sizable beams. It is worth noting that most optimization strategies considered in the available literature dedicated to resource allocation make use of genetic algorithms or neural networks. In [15] the Simulated Annealing (SA) algorithm [16], has been proposed to minimize the co-channel interference in the uplink of two independent satellite systems. In [17] a reinforcement learning approach against satellite channel impairments such as orbital perturbations and atmospheric effects is presented. As a side remark, we point out that the problem of radio resource management (RRM) has been studied also in the context of terrestrial networks [18] [19]. However, the payload constraints and the different network topologies make the two optimization problems significantly different.

The original contribution of the present paper with respect to the state of the art is the following[2]:

- We propose a novel resource allocation strategy in which a multi-objective optimization problem is addressed through the definition of an *ad-hoc* objective function taking both fairness and absolute capacity mismatch into account.

[2]Part of the content of the present paper has been presented at the Advanced Satellite Multimedia Systems (ASMS) conference 2016 [20].

- Analytical insights are presented for such function and it is shown that its convexity can not be guaranteed for all payload constraints.
- Unlike [15], the present paper applies a variant of SA to the forward link of a multibeam satellite system equipped with a payload which is flexible in terms of power allocation, bandwidth allocation or both.
- We perform a detailed analysis of the computational complexity of the proposed optimization algorithm showing that it grows linearly with relevant system parameters such as the number of TWTAs in the payload.
- We show the potential of the proposed resource allocation scheme considering a realistic requested traffic profile, realistic operative conditions (TWTA characteristics, intermodulation interference, beam pattern) and considering the DVB-S2 standard as reference for the physical layer.
- We compare the performance of the new objective function with others previously proposed in literature and show that it achieves comparatively good performance in both low traffic and high traffic regimes.

We stress the fact that the proposed solution is tested under realistic conditions and taking into account effects such as intermodulation interference, ModCod constraints, dependency between symbol constellation and nonlinear effects in the transponder, antenna radiation pattern and consequent inter-beam interference as well as realistic traffic request time variation. Up to our knowledge all such effects have not been jointly considered so far in previous works on multibeam satellite RRM problems.

The remainder of this paper is structured as follows. Section II presents the system model and the formulation of the resource allocation problem. In Section III we introduce and analyze the properties of the objective function. In Section IV our resource allocation strategy is presented and its complexity analyzed. Section V contains the simulation results, while the conclusions are presented in Section VI.

II. Radio Resource Allocation Problem

A. System Model

A multi-beam geostationary satellite system is taken as reference. The satellite generates a geographical footprint subdivided into N_b beams, where each beam i, $i = 1, \ldots, N_b$, serves N_u^i fixed satellite terminals. The population of users active on beam i generates an aggregate traffic request which we denote as T_r^i. Let us denote with $h_{i(j),i'}^j$ the gain experienced by the signal transmitted in beam i' and received by user j located in the footprint of beam $i(j)$, where j can take values between 1 and N_u^i and $N_u^{tot} = \sum_{i=1}^{N_b} N_u^i$, N_u^{tot} being the total number of users in the system. Such gains account for the receiving and transmitting antennas gains and the propagation impairments (e.g., free space loss and atmospheric attenuation). Ideally, each satellite terminal j is expected to receive only the signal transmitted by its reference beam, which we denote as $i(j)$. However, due to the secondary lobes of the satellite antennas, user terminals suffer from interference generated by beams others than the reference one operating in the same frequency band (co-channel beams), leading to $h_{i(j),i'}^j \neq 0$ for some $i' \neq i(j)$. For ease of notation in the following we will indicate the reference beam with i and the interference beams as i'. As far as the payload model is concerned, a single feed per beam (SFPB) architecture is considered. It is assumed that a number of TWTAs equal to N_{TWTA} is available on-board the satellite and that each of them amplifies the same amount of bandwidth. Each tube has a maximum available power equal to P_{tot}. Depending on the operating point of the tube a higher or lower power efficiency[3] can be obtained. Each TWTA reuses the whole bandwidth and serves a subset of beams. The association between beams and the

[3]The power efficiency is intended as the ratio of power transmitted in RF with respect to the DC power provided to the tube.

TWTAs is specified by a so-called *connection matrix*. We consider as a reference, or *conventional* system, one in which the total useful bandwidth[4] B of each TWTA is shared uniformly among the subset of amplified beams, so that the bandwidth per beam depends only on the specific coloring scheme adopted (e.g., B, $B/2$, or $B/4$ for 1, 2, or 4 colours, respectively) and can not be modified in order to follow the traffic request variations. The TWTA bandwidth is shared among the connected beams in a way that such beams cannot have overlapping portions of bandwidth. Data is transmitted through a beam making use of multiple carriers, each being assigned a fraction of the bandwidth allocated to the beam. The portions of data traffic addressed to the users served by a given beam are multiplexed in time according to a time division multiplexing (TDM) framing.

B. Problem Formulation

Our goal is to allocate resources such that each beam receives an offered capacity T_o^i, $i \in \{1, \ldots, N_b\}$, that is as close as possible to the requested capacity T_r^i while taking fairness into account. T_o^i depends on the bandwidth allocated to beam i, the power settings of the TWTA to which beam i is connected, as well as on the co-channel interference generated by other beams and on the channel gains (relative to both reference signal and interferers) of each single user. The resource allocation takes place at the gateway. We assume that the gateway has knowledge of the gains $G_{i,j}$, $i \in \{1, \ldots, N_b\}$, $j \in \{1, \ldots, N_u^i\}$, N_u^i being the number of users served by beam i. This assumption is a realistic one since fixed terminals are considered and thus the rate of channel variation can be assumed to be relatively slow. The channel gains are assumed to be periodically estimated by the gateway through a return channel. This is actually the case in real systems, in which knowledge of such gains are used by the ground station to choose the ModCod which is best suited to each terminal's current channel condition, i.e., the ModCod with the highest spectral efficiency that can be supported by the channel.

We consider three different payloads. The first payload can be optimized both in terms of bandwidth and power allocation. The second one has only bandwidth flexibility while in the third payload only the TWTA operating conditions in terms of power can be adjusted. For a fair comparison, both the number and the characteristics of the TWTAs are the same for all payloads. We also assume that the connection matrix, which determines the subset of beams that are connected to each TWTA, is the same in all payloads. The subset of beams connected to different TWTAs are disjoint, i.e., one beam cannot be connected to more than one TWTA. The flexibility in terms of bandwidth allows to modify, in each TWTA, the spectrum assignment to the subset of beams connected to it, under the constraint that the same portion of spectrum can not be assigned to more than one beam connected to the same TWTA. Note that the same portion of spectrum can be assigned from different TWTAs to one of the beams in their relative subset, so that from a resource allocation point of view each TWTA acts as an independent unit. The bandwidth can be adjusted with a granularity of B_{ch} MHz, i.e., the bandwidth allocated to a beam must be a multiple of B_{ch} MHz. In the following we will refer to such elementary unit of bandwidth as *chunk*. Thus, the amount of bandwidth that can be assigned to a certain beam can be expressed as $N_{\text{ch}} B_{ch}$, where N_{ch} belongs to the set $\{1, 2, \ldots, N_{\text{ch}}^{\text{tot}}\}$, $N_{\text{ch}}^{\text{tot}}$ being the maximum number of chunks available in the system. Transmission takes place on different carriers, each corresponding to a chunk.

The flexibility in terms of power allocation consists in the possibility to change operating point (i.e., input back-off (IBO)) and power setting of each TWTA independently of the others. The power setting defines the attenuation (in dB) of the radio frequency saturated power delivered by the TWTA with respect to the peak in the reference operation mode (no attenuation). A larger power profile indicates larger output back-off (OBO) for a given IBO, which reduces the non-linear effects of the tube at the cost of a reduced power efficiency. The power settings can be chosen from the set $\{0, 1, \ldots, S_p - 1, S_p\}$, where

[4]by useful bandwidth we mean the overall bandwidth once the guard bands between adjacent channels has been subtracted.

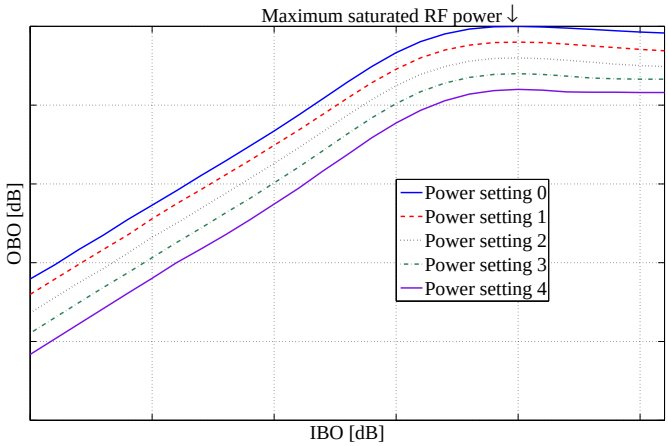

Fig. 1: Example of TWTA characteristics for five different power settings ($S_p = 4$). The operation point corresponding to the maximum saturated RF power is also indicated.

0 corresponds to the reference setting. An example of TWTA characteristics for five different power settings (i.e., $S_p = 4$) is shown in Fig. 1. Payloads with no power flexibility keep a fixed IBO (equal to 3 dB) and a power profile equal to 2.

The optimization problem that we aim to solve is:

$$\begin{aligned} \underset{\boldsymbol{v},\boldsymbol{p},\boldsymbol{B}}{\text{minimize}} \quad & f(\boldsymbol{v},\boldsymbol{p},\boldsymbol{B}) \\ \text{subject to} \quad & v_t \in \{-\text{IBO}_{\text{max}}, -\text{IBO}_{\text{max}} + \Delta\text{IBO}, \ldots, -\text{IBO}_{\text{min}}\} \\ & p_t \in \{0, 1, \ldots, S_p\} \\ & \forall t \in \{1, 2, \ldots, N_{\text{TWTA}}\} \\ & \boldsymbol{B} \in \mathcal{B} \end{aligned}$$

where N_{TWTA} is the number of TWTAs in the payload, $f(\boldsymbol{v},\boldsymbol{p},\boldsymbol{B})$ is the objective function to be minimized, which will be defined Section III, $\boldsymbol{v} = (v_1, \ldots, v_{N_{\text{TWTA}}})$ and $\boldsymbol{p} = (p_1, \ldots, p_{N_{\text{TWTA}}})$ are vectors containing the IBO's and the power profiles for all TWTAs, respectively, while $\boldsymbol{B} \in \mathcal{B} \subset \mathbb{Z}^{N_b \times N_{\text{ch}}}$ is the bandwidth allocation matrix, which belongs to the set of feasible bandwidth allocation matrices $\mathcal{B}$. $\mathcal{B}$ is a subset of the set of binary matrices whose structure depends on the specific payload bandwidth constraints. ΔIBO is the granularity in the IBO setting and is a payload specific parameter. As in real systems, the power optimization of each TWTA is done by selecting the IBO which, in turn, gives a certain OBO.

As a common practice in optimization problems, we aim at minimizing an objective function. The objective function reflects the system key performance indicators (KPI). In order to define the KPIs we start with some general considerations. Our goal is to efficiently allocate the satellite resources with the aim of satisfying the requested traffic in all beams. The capacity request satisfaction can be looked at from a system (or global) perspective as well as from a beam (or user) perspective. From a global perspective, a valid choice would be to take as objective function a measure of the error in matching the requested capacity across the beams, i.e.,

$$\text{E} = \sum_{i=1}^{N_b} (T_o^i - T_r^i)^2.$$

Although this may be a valid indicator to measure the error with respect to an ideal resource allocation condition (by ideal we mean one in which the offered capacity exactly matches the requested capacity in each beam), the offered capacity exceeding the requested one is treated in the same way as the missing capacity, which is not desirable. Moreover, the measure E is potentially unbounded (i.e., it can assume arbitrary positive values) and this makes it difficult to evaluate the goodness of a

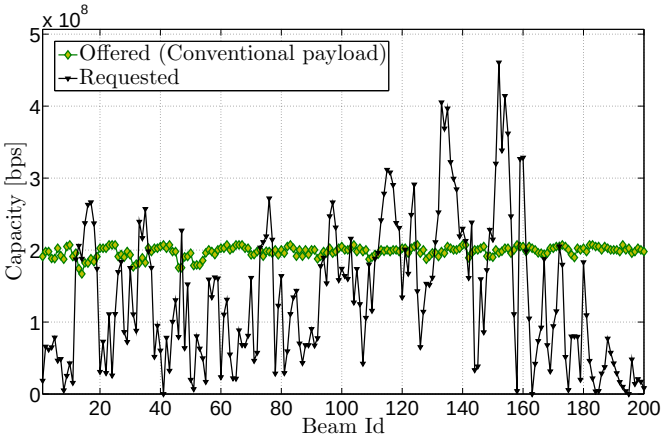

Fig. 2: Example of requested capacity and offered capacity in bits per second (bps) for the conventional payload plotted versus beam Id.

given solution. Furthermore, even if relatively good results are obtained in terms of matching error E, it can still happen that traffic requests are largely unmatched for a non-negligible number of beams. This may be the case mainly for beams that present relatively low requests and for which too little resources are allocated. As a matter of fact, beams with relatively little capacity request may not have great impact on E and thus an optimization solution that performs well at a global scale may neglect such beams. Although beams with higher capacity requests are likely to be the most profitable ones from the satellite operator perspective, low traffic beams should be taken into account in the optimization process since other considerations may make such beams appealing (e.g., presence in the territory, reputation of the operator, etc.). Fairness is indeed a relevant parameter to be accounted for in the RRM optimization. Several different measures of fairness have been proposed in literature, such as the Jain Index and the normalized entropy. Using one of these measures as (negative) objective function would have the disadvantage of not accounting for the absolute value of the mismatch in terms of capacity, either missing or in excess.

Let us now define the satisfaction index of beam i as $\text{SI}_i = \frac{T_o^i}{T_r^i}$. SI is a non-negative number which gives a measure of the extent up to which the requested capacity is satisfied. If SI< 1 the beam has been allocated insufficient resources for its capacity needs, while SI> 1 indicates that the beam is being over-provisioned. Ideally it would be desirable to keep track of both request satisfaction and absolute gap between the requested and the offered capacity. A way to visualize the system state in such terms is to represent of all beams as points of a scatter plot in a plane having as axes the satisfaction index and the difference $\Delta_i = T_o^i - T_r^i$, which gives a measure of the missing (if negative) or exceeded (if positive) capacity. We refer to such plane as the satisfaction/gap (SG) plane. In Fig. 2 an example of requested and offered capacity for the conventional payload plotted versus the beam Id is shown. The corresponding SG representation is depicted in Fig. 3.

III. Objective Function

The representation in the SG plane of the system state in terms of offered/requested capacity gives a qualitative idea of the goodness of a resource allocation solution in terms of both satisfaction and gap distribution. In order to have also a quantitative measure, we define in the following an objective function which is derived from the SG plane. We will refer to it as the satisfaction-gap measure (SGM).

A. SGM Definition

Our aim is to find an objective function that satisfies the following:

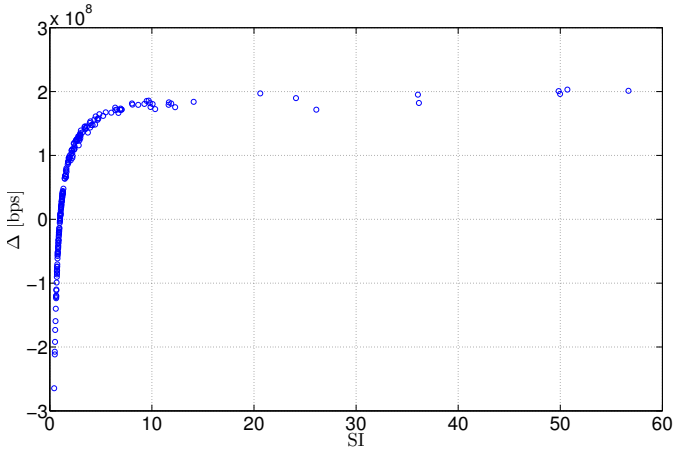

Fig. 3: Representation of beams in the SG plane. Each of the 200 beams is represented as a point (blue circle) in the plane having as x coordinate the satisfaction index and as y coordinate the capacity gap Δ as defined in this section.

1) Provides a measure of the mismatch with respect to the ideal case accounting for both capacity gap and satisfaction in all beams.
2) A beam with satisfaction $1-\delta$, with $0 \leq \delta < 1$, should weight more than a beam with satisfaction $1+\delta$ (which is also undesired but not as bad as having beams with missing capacity).
3) Assumes values in the interval $[0, 1]$, where 1 corresponds to the ideal state in which there is perfect match between offered and requested capacity through all beams[5].

We apply a transformation to the SG plane in such a way that the measure we look for satisfies the three conditions above. Let us start with point 1). In order to take both SI and Δ into account, we treat the SG plane as a complex plane, in which SI represents the real axis and Δ the imaginary axis. Beam i, $i \in \{1, \ldots, N_b\}$, is represented as a complex number in such plane, with coordinates:

$$c_i = \mathrm{SI}_i + j\Delta_i$$

where $j = \sqrt{-1}$. In the following we drop the index i to simplify the notation. In order to satisfy point 2) we apply the following transformation to the beams having real part lower than 1:

$$\mathrm{Re}\{c\} \rightarrow 1 - \frac{1}{\mathrm{Re}\{c\}}, \ \forall c : \mathrm{Re}\{c\} \leq 1. \tag{1}$$

This transformation translates a smaller SI into a larger distance from the origin. In order to satisfy point 3) we shift the points with real part (satisfaction) larger than or equal to 1 towards the origin by applying the transformation $c \rightarrow c - 1$. In this way the point representing the optimal solution becomes $(0, 0)$. In Fig. 3 it can be seen how, depending on the unit of measure adopted to measure the excess/missing capacity (e.g., kbps, Mbps, Gbps) the range of the y axis can be quite wide with respect to the x axis. This can be easily fixed with the following scaling operation:

$$\mathrm{Im}\{c\} \rightarrow \frac{\mathrm{Im}\{c\}}{\beta} \tag{2}$$

with $\beta > 0$. The value of β can be chosen, for instance, equal to (or a function of) the system throughput of the conventional payload. In this way it is possible to make a comparison in terms of the goodness in the resource allocation solution between systems with different total capacities.

[5]Note that this can not be achieved in systems that are under-dimensioned in terms of total system bandwidth and TWTA number.

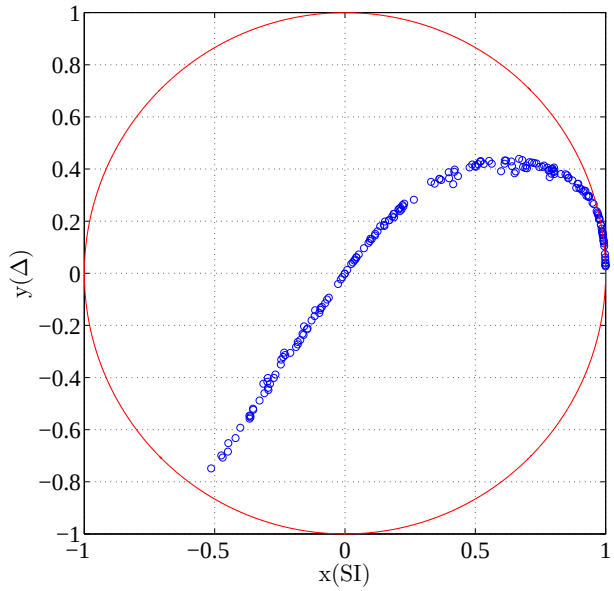

Fig. 4: Representation of beams in the modified SG plane. Each beam is represented as a point (blue circle) in the plane having as x coordinate a function of the satisfaction index and as y coordinate a function of the capacity gap Δ. The unitary circle is also shown (red line).

In order to get a measure which takes values between 0 and 1, we apply one last transformation to the plane which confines all the points within the circle of radius 1 around the origin. This is done applying the following transformation to the absolute value of each point, without modifying its phase:

$$\mid c \mid \rightarrow 1 - e^{-|c|}. \tag{3}$$

According to this transformation, a point at infinite distance from the ideal condition (origin) will lie on the unitary circle, while a point that has $\mid c \mid \ll 1$ before applying (3) will have a distance from the origin approximately equal to $\mid c \mid$. The modified plot corresponding to the example in Fig. 3 is shown in Fig. 4.

Starting from the transformed plot, we define the SGM as a measure of the average distance from the optimal condition:

$$\text{SGM} = 1 - \frac{1}{N_b} \sum_{i=1}^{N_b} \mid c_i \mid^3 . \tag{4}$$

The third power rise in the sum on the right hand side of expression (4) is included in order to give more weight to beams that are farther apart from the ideal condition, i.e., have SI which is either close to zero or much larger than 1 or have a large mismatch in terms of absolute capacity. Larger values have been tested but lead no significant advantage.

The SGM is the complement to 1 of the average (cube of the) distance from the origin of the points in the transformed scatter plot. It can be easily seen that such measure takes values in $[0, 1]$ and is close to 1 when all points are gathered around the origin, which corresponds to the case in which the offered capacity matches almost exactly the requested capacity in each of the beams and the Δ's are relatively small.

The optimization problem to be solved is, finally:

$$\begin{aligned} \underset{\boldsymbol{v},\boldsymbol{p},\boldsymbol{B}}{\text{minimize}} \quad & -\text{SGM}(\boldsymbol{v},\boldsymbol{p},\boldsymbol{B}) \\ \text{subject to} \quad & v_t \in \{-\text{IBO}_{\text{max}}, -\text{IBO}_{\text{max}}+1, \ldots, -\text{IBO}_{\text{min}}\} \\ & p_t \in \{0,1,\ldots,S_p\} \\ & \forall t \in \{1,2,\ldots,N_{\text{TWTA}}\} \\ & \boldsymbol{B} \in \mathcal{B}. \end{aligned}$$

B. SGM Analysis

In the following we provide some analytical insight on the SGM.

We start by considering the implications of the transformation undergone by the points in the SG plane according to expressions (1)-(3). Following transformation (1), the points in the SG plane corresponding to beams with SI≤ 1 are modified as follows:

$$c \to 1 - \frac{1}{\text{SI}} + j\Delta. \tag{5}$$

Transformation (2) is a simple re-scaling which is applied to all points. The scaling factor β can be chosen at will. One option is to choose β equal to the average requested traffic across all beams, i.e.,

$$\beta = \frac{1}{N_b}\sum_{i=1}^{N_b} T_r^i.$$

After applying transformation (3) the absolute value of each point in the SG plane is transformed as follows:

$$|c| \to \begin{cases} 1 - \exp\left(-\sqrt{\left(1-\frac{1}{\text{SI}}\right)^2 + \frac{\Delta^2}{\beta^2}}\right), & \text{if SI} \leq 1 \\ 1 - \exp\left(-\sqrt{\text{SI}^2 + \frac{\Delta^2}{\beta^2}}\right), & \text{if SI} > 1. \end{cases} \tag{6}$$

Plugging Eqn. (6) into Eqn. (4) we have:

$$\begin{aligned} \text{SGM} &= 1 - \frac{1}{N_b}\sum_{i|\text{SI}_i\leq 1} |\, c_i \,|^3 - \frac{1}{N_b}\sum_{i|\text{SI}_i>1} |\, c_i \,|^3 \\ &= 1 - \frac{1}{N_b}\sum_{i|\text{SI}_i\leq 1} \left|1 - e^{-\theta_i^u}\right|^3 - \frac{1}{N_b}\sum_{i|\text{SI}_i>1} \left|1 - e^{-\theta_i^o}\right|^3 \\ &= \frac{3}{N_b}\sum_{i|\text{SI}_i\leq 1} e^{-\theta_i^u}\left(1 - e^{-\theta_i^u} + \frac{e^{-2\theta_i^u}}{3}\right) \\ &+ \frac{3}{N_b}\sum_{i|\text{SI}_i>1} e^{-\theta_i^o}\left(1 - e^{-\theta_i^o} + \frac{e^{-2\theta_i^o}}{3}\right) \end{aligned} \tag{7}$$

where we defined $\theta_i^u = \sqrt{\left(1-\frac{1}{\text{SI}_i}\right)^2 + \frac{\Delta_i^2}{\beta^2}}$ and $\theta_i^o = \sqrt{\text{SI}_i^2 + \frac{\Delta_i^2}{\beta^2}}$ as the argument of the exponential term for underprovisioned and overprovisioned beams, respectively. If the system under consideration is characterized by a significant mismatch in terms of SI or Δ in all the beams, expression (7) can be simplified by neglecting the terms of higher order. In this case we have:

$$\text{SGM} \simeq \frac{3}{N_b}\sum_{i|\text{SI}_i\leq 1} e^{-\theta_i^u} + \frac{3}{N_b}\sum_{i|\text{SI}_i>1} e^{-\theta_i^o}. \tag{8}$$

From Eqn. (8) some interesting conclusions can be drawn about the way the two performance measures combined in the SGM affect the objective function. Let us now focus on beams in the following situations: one in which there is a significant

mismatch in terms of absolute capacity while having a satisfaction index close to 1 (Δ-mismatch) and a second situation in which the absolute capacity mismatch is small while the satisfaction index mismatch is much larger than one (SI-mismatch).

1) Δ-Mismatch: In this case the mismatch in terms of SI is much smaller than the capacity mismatch, specifically $\frac{\Delta_i^2}{\beta^2} \gg \left(1-\frac{1}{\mathrm{SI}}\right)^2$ for the underprovisioned beams and $\frac{\Delta_i^2}{\beta^2} \gg \mathrm{SI}^2$ for the overprovisioned beams. For underprovisioned beams, this represents a situation in which a beam with a capacity request much larger than the average has an offered capacity which is slightly smaller in percentage to the requested one [6]. Similar considerations can be done for overprovisioned beams. In this case the terms in the sums of Eqn. (7) corresponding to beams with Δ-mismatch can be approximated as:

$$\begin{cases} e^{-\theta_i^u} = e^{-\sqrt{\left(1-\frac{1}{\mathrm{SI}}\right)^2+\frac{\Delta_i^2}{\beta^2}}} \simeq e^{-\frac{|\Delta|}{\beta}}, & \text{for SI} \leq 1 \\ e^{-\theta_i^o} = e^{-\sqrt{\mathrm{SI}^2+\frac{\Delta_i^2}{\beta^2}}} \simeq e^{-\frac{|\Delta|}{\beta}} & \text{for SI} > 1. \end{cases} \tag{9}$$

2) SI-Mismatch: In case of SI-mismatch, a beam is not in an ideal condition mostly due to the SI axis in the SG plane. Formally, we say that a beam has an SI-mismatch if $\left(1-\frac{1}{\mathrm{SI}}\right)^2 \gg \frac{\Delta_i^2}{\beta^2}$ in case of underprovisioned beam, or $\mathrm{SI} \gg \frac{\Delta_i^2}{\beta^2}$ for an overprovisioned beam. The terms in the sums of Eqn. (7) corresponding to beams with SI-mismatch can be approximated as:

$$\begin{cases} e^{-\theta_i^u} = e^{-\sqrt{\left(1-\frac{1}{\mathrm{SI}}\right)^2+\frac{\Delta_i^2}{\beta^2}}} \simeq e^{-\left|1-\frac{1}{\mathrm{SI}}\right|} & \text{for SI} \leq 1 \\ e^{-\theta_i^o} = e^{-\sqrt{\mathrm{SI}^2+\frac{\Delta_i^2}{\beta^2}}} \simeq e^{-\mathrm{SI}} & \text{for SI} > 1. \end{cases} \tag{10}$$

In both mismatch cases we see that, in a first order approximation, the weight of a beam on the overall SGM calculation depends exponentially on the metric for which the mismatch is larger. In intermediate cases the weight of a beam in the SGM depends on a combination of the two metrics.

Now we show that for some payload constraints the objective function is non-convex. Since the objective function is defined as $f = -\mathrm{SGM}$, this is equivalent to prove the non-concavity of SGM. Note that for some specific systems characterized by a given channel matrix and traffic request and for certain power and bandwidth allocation constraints the function may be convex. However, since the user population as well as the traffic requests may change over time, the convexity of $-\mathrm{SGM}$ may not always hold. This is shown in the following proposition.

Proposition 3.1: There exist some system configurations for which the SGM function is non concave.

Proof 3.2: See Appendix.

Since an algorithm designed to work well for a convex function may perform poorly if the function is not convex, we opt for a stochastic optimization solution, which is presented in the next section. Considerations on the performance of such algorithm in the convex case are also presented.

IV. Resource Allocation Strategy

Even assuming full channel state information at the transmitter, finding the optimal resource allocation is not trivial. This is due, on one side, to the non-convexity of the objective function and on the other side to the large number of possible solutions, which makes exhaustive search not viable [7].

[6]For instance, consider a system with average requested capacity equal to 100 Mbps. Assuming a beam with a requested capacity equal to 900 Mbps and an SI equal to 0.93, we have $\frac{\Delta_i^2}{\beta^2} \simeq 0.397$ while $\left(1-\frac{1}{\mathrm{SI}}\right)^2 = 0.0057$.

[7]If we consider a flexible payload with 50 TWTAs, a number of bandwidth chunks equal to 8 and a number of allowed IBO levels equal to 10, the number of possible allocations (feasible points) is equal to $(256 \times 10)^{50}$ which is on the order of 10^{170}.

We propose a suboptimal algorithm based on a slightly modified version of the Simulated Annealing algorithm [16]. The algorithm tries to minimize the objective function defined in Section II-B. This is done by running iteratively the SA algorithm, each time using lower starting and stopping temperatures. The way the SA algorithm is applied at each run is described in the following.

A. Perturbation of the Feasible Point

The SA algorithm uses as starting point the same bandwidth and power allocation as the conventional payload.

At each iteration the algorithm perturbates the feasible point. Depending on the payload to which the algorithm is applied, either the bandwidth, the power or both can be modified. For a payload with full flexibility the algorithm chooses randomly at each iteration whether to modify one or the other.

The perturbation of the feasible point is done as follows. A beam is selected at random, then:

- If the bandwidth is to be modified, the number of bandwidth chunks N_{ch} currently allocated to the beam is modified by adding to such number a random variable $u \in \{-1, 0, +1\}$ while keeping the number of allocated beams within the set $\{1, 2, \ldots, N_{\text{ch}}^{tot} - 1, N_{\text{ch}}^{tot}\}$. Once the new number of chunks N_{ch} is fixed, their spectral location within the available bandwidth is selected at random among the $\binom{N_{\text{ch}}^{tot}}{N_{\text{ch}}}$ possible dispositions. Afterwards, the algorithms switches off the chunks allocated to the selected beam from the other beams connected to the same TWTA (if necessary).
- If the power is to be modified, the algorithm selects the TWTA to which the selected beam is connected and modifies either its IBO or its power profile. Modifying the operating conditions of the TWTA induces a modification in the amount of power delivered by the TWTA, its power efficiency (i.e., the ratio of the delivered RF power to the absorbed DC power) and the intermodulation interference associated with the TWTA nonlinearity. All these effects are taken into account by the algorithm through a realistic payload model. Note also that all the beams connected to the same TWTA are affected by the same attenuation/amplification of the signal on the selected beam. This is done in order to avoid the so-called *capture effect*, which takes place in TWTAs when carriers of different power are fed to the amplifier [21].

B. SGM Evaluation

Once the resources of the TWTA corresponding to the selected beam have been modified, the resulting signal-to-interference-plus-noise ratio (SINR) of all users are calculated for each bandwidth chunk[8]. The new SINR's are used to determine the ModCod with highest spectral efficiency supported by the channel of each terminal in each chunk according to the system specifications. Once the spectral efficiencies for all terminals are obtained, they are averaged out across users and chunks. We call $\eta_{i,c}^{j}(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{B})$ the spectral efficiency achievable by terminal j in chunk c of beam i at the feasible point $(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{B})$. In most works dedicated to RRM optimization the spectral efficiency is evaluated using the Shannon capacity formula. Although such approach is useful to get an insight on the theoretical performance limits of the system, the transmission rate of real systems is constrained by the symbol constellation size as well as by the finiteness of the channel codeword lengths. In current satellite systems it is common practice to use look-up tables that contain the highest ModCod (and thus the spectral efficiency) achievable at a given SINR. The SINR seen by terminal j in chunk c of beam i at the feasible point $(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{B})$ is:

$$\text{SINR}_{i,c}^{j}(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{B}) = \frac{P(v_{t(i)}, p_{t(i)})|h_{i,i}^{j}|^2 b_{i,c}}{\sum_{i',i' \neq i} P(v_{t(i')}, p_{t(i')})|h_{i,i'}^{j}|^2 b_{i',c} + N} \tag{11}$$

[8]The SINR is calculated on a chunk-by-chunk basis since each chunk is assumed to be a single carrier.

where we indicated with $h^j_{i,i'}$ the channel coefficient from beam i' to user j in beam i. $h^j_{i,i'}$ accounts for the propagation loss as well as the satellite and the terminal antenna gains, $t(i)$ identifies the TWTA feeding beam i while N represents the AWGN noise power. $b_{i',c}$ represents element (i', c) in the binary matrix $\boldsymbol{B}$, and takes value 1 or 0 depending on whether chunk c is allocated to beam i' or not, respectively. According to this notation, $P(v_{t(i)}, p_{t(i)})$ represents the power allocated to beam i, which is delivered by TWTA $t(i)$, in correspondence of the IBO $v_{t(i)}$ and the TWTA power profile $p_{t(i)}$. Finally, N represents the AWGN noise power which is assumed to be the same for all the terminals. The average spectral efficiency in beam i is then:

$$\overline{\eta}_i = \frac{1}{N^i_u N^i_{\text{ch}}} \sum_{j=1}^{N^i_u} \sum_{c=1}^{N^i_{\text{ch}}} \eta^j_{i,c}(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{B}) \tag{12}$$

where N^i_u is the number of users in beam i while N^i_{ch} is the number of chunks allocated to beam i. Expression (12) follows from the assumption that all users within a beam access their content in a TDM fashion such that each user is allocated the whole chunk of bandwidth during the assigned reception slot. Using the average spectral efficiency together with the number of bandwidth chunks allocated to each beam and taking the roll-off into account, the new value of the offered capacity is calculated as follows:

$$T^i_o = \frac{N^i_{\text{ch}} B_{ch}}{1+\alpha} \overline{\eta}_i \tag{13}$$

α being the roll-off factor. Finally, using Eqn. (13), the SGM is calculated as described in Section III.

C. Feasible Point Update

Once the new SGM is calculated, there are two possibilities:

- The SGM obtained in the new point is larger than the old one. In this case the new point is kept and a new iteration starts.
- The SGM obtained in the new point is smaller than the old one. In this case the new point is kept with a certain probability. The probability of keeping the new point depends on a simulation parameter that is updated periodically. In the literature related to simulation annealing, such parameter is called *temperature* and is indicated in the following as T_{sa}. The new point is accepted with probability:
$$\exp\left(-\frac{\text{SGM}_{\text{old}} - \text{SGM}_{\text{new}}}{\text{SGM}_{\text{old}} T_{\text{sa}}}\right).$$
Since we are considering the case in which the new point is worse than the previous one, in the expression above we have $\text{SGM}_{old} - \text{SGM}_{new} > 0$. Note also that such probability decreases as the temperature T_{sa} decreases. Similar expressions are commonly found in literature, except for division by SGM_{old}, which we use to increase the sensitivity to SGM variations.

The temperature is decreased once a predefined number of iterations is reached. The cooling law at iteration n is:

$$T_{\text{sa}}(n) = \Delta T \times T_{\text{sa}}(n-1) \tag{14}$$

where $0 < \Delta T < 1$ is a design parameter. The block diagram describing one call of the algorithm is shown in Fig. 5.

As mentioned earlier in this section, the proposed algorithm is a modified version of the SA. The modification consists in that the SA algorithm is run iteratively, each time using lower starting and stopping temperatures. Specifically, if we indicate with `T_start`[l] and `T_stop`[l] the starting and stopping temperatures at the l-th algorithm call, respectively, the following holds:

$$\texttt{T\_start}[l] = \texttt{T\_stop}[l-1].$$

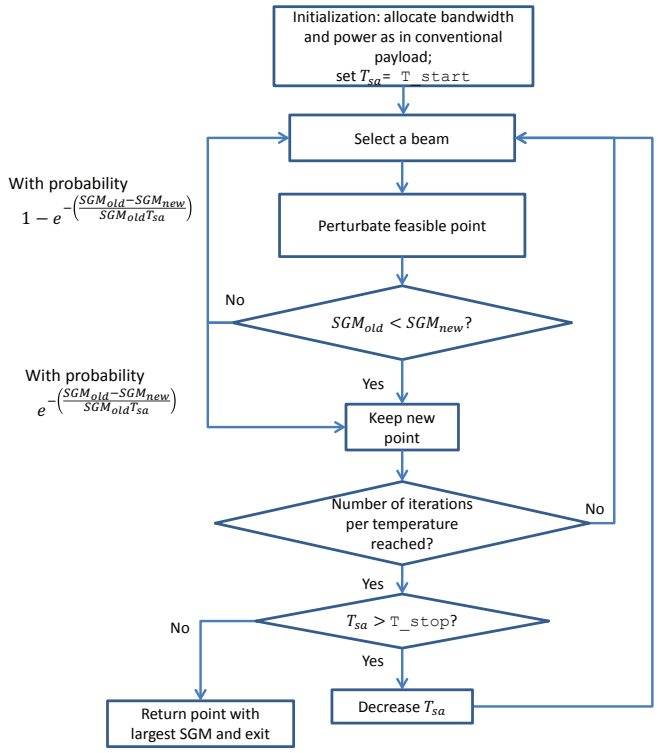

Fig. 5: Flow diagram for one call of the SA algorithm. At each call the initial and starting temperatures T_start and T_stop are decreased such that T_start[l] = T_stop[$l-1$], l being the call index.

The reason behind such modification is that, for the specific problem considered, we observed a tendency of the SA to converge to local minima. This is a well known behavior of stochastic optimization algorithm with non convex objective functions. The solution usually adopted is to run the algorithm more than once, each time starting from a different starting point. In the setup considered in the present paper, such solution showed limited advantages. For this reason we introduced a variant of such approach in which i) the starting point of the new run is the feasible point output of the previous one, ii) rather than running the SA anew, we decrease the starting and stopping temperature at each call. The overall effect is to break the path of the algorithm in the feasible set, avoiding that it gets stuck in regions characterized by values of SGM that are lower than the last accepted one.

As mentioned in Section III, the choice of a stochastic algorithm is due to the fact that the objective function can be non-convex depending on the specific system. There is a large body of literature indicating that SA has good performance when applied to non-convex problems. Even if the problem is not convex, however, the SA algorithm has been shown to lead to good results. See [22] and references therein for further details.

D. Analysis of Computational Complexity

The SA algorithm can be seen as a hybrid between a random walk and a greedy algorithm. The larger the temperature T_{sa}, the closer the behavior of the algorithm is to that of a random walk. As shown in Fig. 5, at each iteration of the algorithm a random perturbation of the feasible point is performed followed by the evaluation of the new value of the SGM and a comparison with the old one. Within an iteration, the most demanding part in terms of number of elementary operations is the evaluation of the feasible point. This requires the calculation of the SINR for each user in the network on each chunk of bandwidth as shown in Eqn. (11). The sum in the denominator of (11) has at most N_{TWTA} terms since each bandwidth chunk can be used at most once by each TWTA. For each term in the sum, a multiplication between real numbers is performed.

In order to get an estimation of the complexity growth as a function of the main system and algorithm parameters, in the following we assume that a multiplication or a division between two m bits numbers requires m^2 elementary operations while m operations are required for an addition or a subtraction [9]. Assuming that each number is represented with an m-bits precision, $N_{\text{TWTA}} \times (m^2 + m)$ operations are required to calculate the denominator and $2m^2$ more are needed to calculate the numerator and perform the division. The complexity deriving from the calculation of expressions (12) and (13) can be upper bounded as $(N_u^i \times N_{\text{ch}}) \times [N_{\text{TWTA}} \times (m + m^2) + 2m^2] + (N_u^i \times N_{\text{ch}} - 1) \times m + 5m^2$. The total number of operations Ψ_{oc} required to calculate the offered capacity in all the N_b beams of the system is:

$$\begin{aligned}\Psi_{\text{oc}} = \sum_{i=1}^{N_b} N_u^i \times N_{\text{ch}} \times (N_{\text{TWTA}} \times (m^2 + m) + 2m^2) \\ +(N_u^i \times N_{\text{ch}} - 1)m + 5m^2 \\ = N_u^{\text{tot}} \times N_{\text{ch}} \times (N_{\text{TWTA}} \times (m^2 + m) + 2m^2) \\ +(N_u^{\text{tot}} \times N_{\text{ch}} - N_b)m + 5N_b m^2. \end{aligned} \tag{15}$$

The evaluation of the SGM requires the calculation of N_b terms of the kind shown in expression (6). Lookup tables can be used to evaluate the exponential function, while the calculation of its argument involves two divisions (calculation of SI and division by β^2) three additions (calculation of Δ, sum inside the square root, sum of 1 and the inverse of SI) two multiplications (squares calculation) and a square root, for an overall number of operations that is on the order of $5m^2$+$3m$. The evaluation of Eqn. (4) requires the calculation of N_b cubes, N_b additions and one division, for a total of $(2N_b + 1)m^2 + N_b m$ operations. The overall number of operations Ψ_{iter} required for one iteration of the algorithm can be approximated as:

$$\begin{aligned}\Psi_{\text{iter}} =& N_u^{\text{tot}} \times N_{\text{ch}} \times (N_{\text{TWTA}} \times (m^2 + m) + 2m^2) \\ &+(N_u^{\text{tot}} \times N_{\text{ch}} - N_b)m + N_b(5m^2 + m) + 6m^2 + 3m \\ \approx& N_u^{\text{tot}} \times N_{\text{ch}} \times N_{\text{TWTA}} \times (m^2 + m) \end{aligned} \tag{16}$$

where the last approximation is tight if $N_{\text{TWTA}} \gg 2$ and $N_u^{\text{tot}} \gg N_b$, which is often the case in practical systems. Thus, the evaluation of the SGM requires a complexity which grows as the product of the total number of users in the system, the number of bandwidth chunks and the number of TWTAs. Among such terms, the largest one is usually the number of users in the system.

The SGM is evaluated at each iteration of the algorithm. The total number of iterations within a run depends on the cooling factor (ΔT) and on the value of the initial and final temperature. Specifically, from Eqn. (14) it follows that:

$$T_{\text{sa}}(n) = (\Delta T)^n \, \texttt{T\_start}$$

which implies that the total number n_{tot} of cooling steps to be performed before the algorithm stops is:

$$n_{tot} = \left\lceil \log_{\Delta T} \left(\frac{\texttt{T\_stop}}{\texttt{T\_start}} \right) \right\rceil$$

[9]This is of course an upper bound, since much more efficient methods allow to reduce significantly the number of elementary operations to perform multiplications or additions [23], but the aim of this subsection is to get an estimation of the computational effort needed.

where $\lceil x \rceil$ is the smallest integer larger than or equal to x. For each temperature a certain number of iterations is to be performed. Let us refer to such number as N_{it}[10]. The overall complexity Ψ_{tot} of one run of the algorithm is thus:

$$\Psi_{\text{tot}} \approx N_u^{\text{tot}} \times N_{\text{ch}} \times N_{\text{TWTA}} \times (m^2 + m) \times n_{\text{tot}} \times N_{\text{it}}. \quad (17)$$

The time needed to perform the Ψ_{tot} operations depend on the specific hardware considered. The simulations in the next section took roughly four hours on a commercial laptop with two 2.7 GHz cores and 8 Gigabytes of RAM. Note that one of the most computational demanding parts is the calculation of the SINR for each of the users. Its impact on the complexity can be drastically reduced by implementing the calculation on graphics cards, that can achieve high degree of parallelization and thus notably decrease the computation time. Performance can be further enhanced by using a dedicated hardware. Furthemore, notice that the algorithm does not need to run in real-time. More specifically, it should be run either when some users experience a significant change in their SINR or when the requested traffic changes significantly. Since in the present paper we consider fixed satellite services in the Ka frequency band and above, SINR variations are mostly due to atmospheric events such as rain ant thus the rate of variation is in the order of the tens of minutes or even hours. Furthermore, the rate of traffic variation is usually slow, as shown in the next section, with the requested traffic showing little variations on "coherence" periods of about one hour. Such time correlation can be exploited as follows. Let us assume that a solution to the RRM problem is provided by the algorithm, the next solution can then be calculated using the previous one as a starting point. We observed in our simulations that this can be exploited to speed-up the convergence of the algorithm.

V. Performance Analysis

In the following we present the results obtained by applying the proposed solution to three different payloads. The first payload we consider has full flexibility, in the sense that it can be optimized both in terms of bandwidth and power allocation. The second payload has only bandwidth flexibility, while in the third one only the power settings can be modified. The starting point of the algorithm is the resource allocation used in the conventional payload. In all simulations we fixed $N_{\text{TWTA}} = 50$, an overall useful bandwidth $B = 500$ MHz, bandwidth chunks of $B_{\text{ch}} = 31.25$ MHz (and thus $N_{\text{ch}}^{\text{tot}} = 16$), an IBO granularity $\Delta\text{IBO} = 1$ dB, $S_p = 4$ for the power profiles, $N_b = 200$ beams and $N_u^{\text{tot}} = 2000$ user (with $N_u^i = 10$, $\forall i = 1, \ldots, N_b$). The TWTA characteristics shown in Fig. 1 have been used and the corresponding nonlinear effects have been taken into account. Specifically, an increase in the noise level due to intermodulation has been included in the simulator. Such increase in the noise level depends on the modulation order (and specifically on the amplitude levels in the constellation) and has been accurately taken into account according to the approach described in [24]. The gains of the reference beam and of the interference beams as seen by each user (coefficients $h_{i,i}^j$ and $h_{i,i'}^j$ in Eqn. (11), respectively) have been generated using GRASP, a commercial software for satellite antenna design and analysis [25]. The antenna gain in dB scale as a function of the geographical location of the user terminals is shown in Fig. 6. The average SNR (including the intermodulation interference in the noise calculation but not the interbeam interference) across the users for the conventional payload is 11.6 dB. The bandwidth and power allocation of the conventional payload are used as a starting point for the optimization algorithm.

We compare the results for the different payloads in terms of both the effectiveness of the algorithm to meet the requested capacity and the fairness with which the different beams are treated. Specifically, the Jain Index of the ceiled satisfaction index

[10]It is common practice to set a maximum number of iterations per temperature together with additional stopping conditions that trigger the cooling event before N_{it} iterations are performed. N_{it} can thus be regarded to as an upper bound on the number of iterations per temperature.

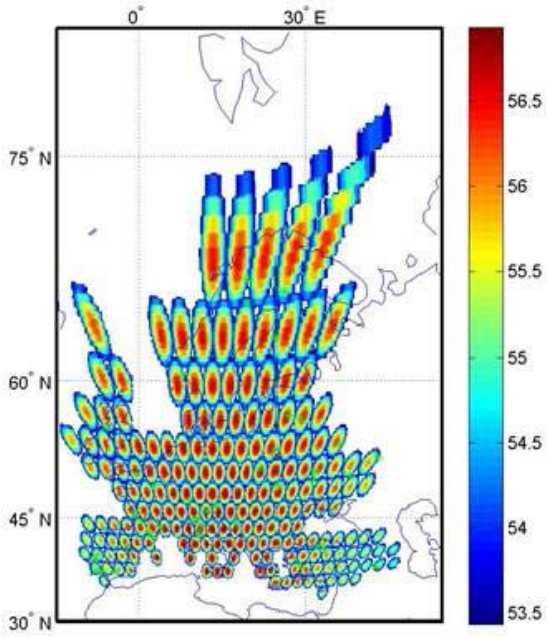

Fig. 6: Antenna gain in dB scale as a function of the geographical location of the user terminals.

is used to measure the fairness in the system. The ceiled satisfaction index is defined as $\overline{\text{SI}} = \min\{\text{SI}, 1\}$ and is a measure of the satisfaction level of a beam which focuses on the missing capacity. The JI is calculated as:

$$\text{JI} = \frac{\left(\sum_{i=1}^{N_b} \overline{\text{SI}}_i\right)^2}{N_b \sum_{i=1}^{N_b} \overline{\text{SI}}_i^2}. \tag{18}$$

Another relevant figure of merit for satellite communications systems is the *unmet capacity* (UC), which is the overall amount of requested capacity that can not be provided. UC is defined as:

$$\text{UC} = \sum_{i=1}^{N_b} \left(T_r^i - T_o^i\right)^+ \tag{19}$$

where $(x)^+ = \max(x, 0)$. Similarly as for UC, we define the *excess capacity* (EC) as:

$$\text{EC} = \sum_{i=1}^{N_b} \left(T_o^i - T_r^i\right)^+ \tag{20}$$

which corresponds to the sum across the beams of the offered capacity which exceeds the requested capacity. The UC and the EC give an indication of the effectiveness of the resources allocation in the system. Finally, we define the *total offered capacity* (TOC) as:

$$\text{TOC} = \sum_{i=1}^{N_b} T_o^i. \tag{21}$$

The requested capacity per beam considered in the following has been generated according to a traffic model developed by the German Aerospace Center (DLR), accounting for the geographical and time variations of traffic requests as well as the availability of the satellite network. The model provides good matches with real requested traffic statistics, as discussed in [6, section III-E], to which the interested reader can refer for more details. The average across the beams of the traffic pattern time evolution obtained with such model is shown in Fig. 7.

The proposed solution has been tested with both static and time varying capacity requests. The results are presented in subsections V-A and V-B, respectively.

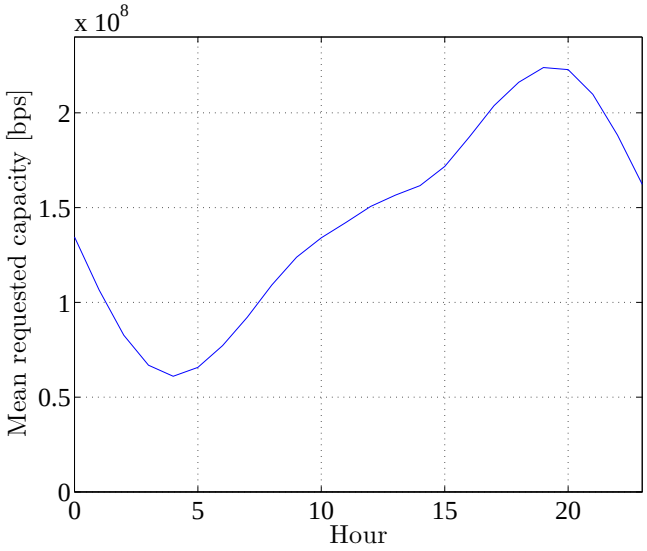

Fig. 7: Spatial average of the requested capacity during a day.

A. Static Capacity Request

In figures 8 and 9 the requested capacity is plotted against the beam Id together with the offered capacity obtained by applying the proposed algorithm to different payloads. Specifically, the capacity offered by the payload with both bandwidth

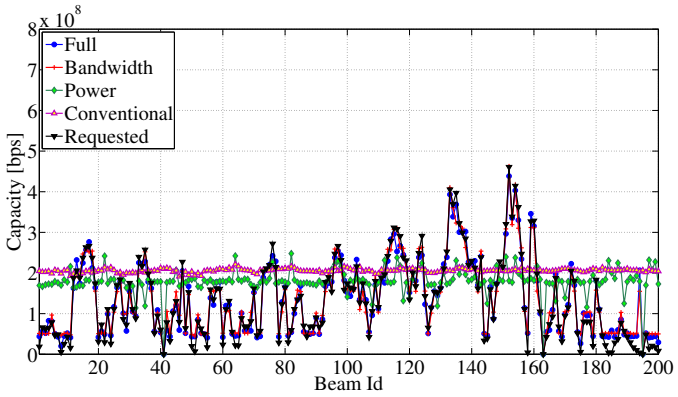

Fig. 8: Capacity versus beam number. The requested capacity at 00:00 h is shown together with the capacity offered by the fully flexible payload, the bandwidth flexible payload, the power flexible payload and the conventional payload.

and power flexibility (Full), the one with only bandwidth flexibility (Bandwidth) and the one with power flexibility only (Power) obtained with the proposed algorithm are shown. The offered capacity of the conventional payload is shown as a benchmark. In Table I and Table II the comparison among the four different payloads is presented for the requested traffic at off peak (00:00) and peak (19:00) hours, respectively. In order to have a deeper understanding of the SGM as performance metric and the implications of using it as objective function, in the tables SGM, Jain Index, unmet capacity and excess capacity are shown for each payload. All values are rounded to the third decimal. Since the last three parameters have been previously used in literature or have an intuitive interpretation, they help to understand the SGM more in depth.

With reference to Fig. 8, the payload with bandwidth flexibility and the payload with full flexibility are able to provide an offered capacity which closely follows the requested one in most of the beams. In beams with very low requested capacity, such as the beam with Id 194, the offered capacity is relatively larger than the requested one. This is in part due to the limited

TABLE I: SGM, Jain Index, unmet capacity and excess capacity at off-peak hour (00:00) for the four payloads. Values are rounded to the third decimal.

	SGM	JI	UC [Gbps]	EC [Gbps]	TOC [Gbps]
Full	0.923	0.995	1.34	1.664	27.222
Bandw.	0.905	0.995	1.33	1.984	27.552
Power	0.62	0.978	4.263	13.26	35.895
Conv.	0.567	0.982	3.237	17.509	39.529

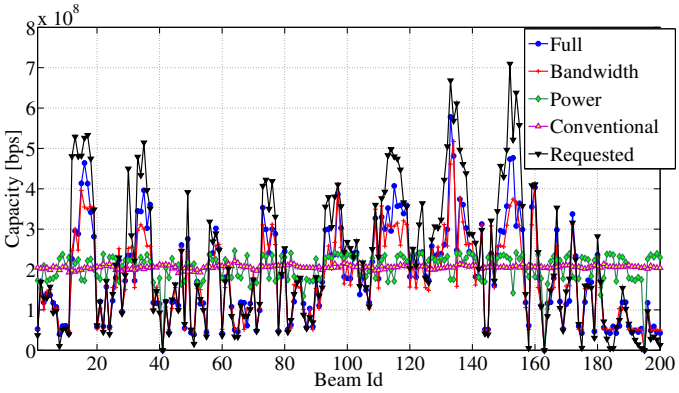

Fig. 9: Capacity versus beam number. The requested capacity at 19:00 h is shown together with the capacity offered by the fully flexible payload, the bandwidth flexible payload, the power flexible payload and the conventional payload.

granularity in terms of bandwidth. The payload with power flexibility is not able to follow the requested traffic as closely as the other two flexible payloads. One of the reasons for this is the fact that the power can be optimized only at TWTA level, so that all beams experience the same power increase or decrease, while this is not the case for the payload with bandwidth flexibility, for which the number of bandwidth chunks assigned to a beam can be different from that of other beams connected to the same TWTA (provided the same chunk of bandwidth is not allocated to more than one of the beams connected to it). As a last remark, we notice that the offered capacity of the conventional payload shows some fluctuations across the beams. These are due to the slight differences in channel gains and interference levels experienced by the different users, that, on turn, depend on the realistic satellite antenna radiation pattern used. From Table I we can see that the qualitative considerations presented above are backed up by the numerical values in the table. As a matter of fact, the SGM is higher in the payloads with full and bandwidth flexibility. This corresponds to higher JI, lower UC and lower EC with respect to the other two payloads. Comparing the Full and the Bandwidth payloads we see that the UC in the two payloads are almost the same while the smallest EC is achieved by the Full payload, which leads to a larger SGM. This shows how the SGM jointly accounts for UC, EC and fairness, although the mapping from SGM to the three measures is not straightforward. It is interesting to note that, as shown in Table I, the Power and the Conventional payloads provide a larger total offered capacity than the other two schemes. Despite of this, the UC is higher in the two systems, which indicates that the Bandwidth and Full payloads make a more efficient use of the satellite resources.

In Fig. 9 we show the results relative to a more demanding pattern of requested capacity. The setup is much more challenging than the one presented in Fig. 8 from an optimization perspective. This can be inferred from the peaks of requested traffic reaching more than three times the capacity offered by the conventional payload, and from the fact that the overall system bandwidth and TWTA number in all the advanced payloads is the same as for the conventional one in both simulations. Also in this case the payloads with full and bandwidth flexibility can best follow the requested traffic profile, while the payload

with power flexibility has only a limited adaptation capability. Interestingly, even though the power flexibility alone is not able to follow the requested traffic, it provides an advantage to the payload with full flexibility. This can be seen from the fact that the Full payload can better approximate the highest peaks with respect to the bandwidth flexible one. This is further confirmed by the results shown in the Table II. In the table we see that the payload with full flexibility achieves the best performance in

TABLE II: SGM, Jain Index, unmet capacity and excess capacity at peak hour (19:00) for the four payloads. Values are rounded to the third decimal.

	SGM	JI	UC [Gbps]	EC [Gbps]	TOC [Gbps]
Full	0.912	0.978	8.514	1.161	37.415
Bandw.	0.884	0.966	10.692	1.081	35.157
Power	0.638	0.93	14.808	10.779	40.738
Conv.	0.603	0.914	15.765	12.166	39.529

all the four figures of merit considered, reducing the unmet and the excess capacity and increasing the system fairness with respect to any of the other payloads. In this case we see how a higher SGM corresponds to a better performance through the whole spectrum of figures of merits considered.

In order to have a better understanding of the benefits deriving from the adoption of the SGM rather than other related objective functions, we applied the proposed optimization algorithm to two different objective functions. One is $-$JI (the "$-$" sign is used so that the optimization problem can be cast as a minimization) while the other is the average absolute value of the capacity gap $\overline{\Delta} \triangleq 1/N_b \sum_{i=1}^{N_b} |\Delta_i|$. We chose such functions because they are the performance metrics at the basis of the SGM definition. The comparison in terms of JI, UC and EC is presented in Table III and Table IV for the off-peak and peak hours, respectively. From Table III we see that applying the proposed algorithm to maximize the Jain index leads to relatively

TABLE III: SGM, Jain Index, unmet capacity and excess capacity at off-peak hour (00:00) for the payload with full flexibility for three different objective functions. Values are rounded to the third decimal.

	JI	UC [Gbps]	EC [Gbps]
SGM, Full	0.995	1.34	1.664
JI, Full	1	0.264	5.048
$\overline{\Delta}$, Full	0.865	1.144	0.738

TABLE IV: SGM, Jain Index, unmet capacity and excess capacity at peak hour (19:00) for the four payloads. Values are rounded to the third decimal.

	JI	UC [Gbps]	EC [Gbps]
SGM, Full	0.978	8.514	1.161
JI, Full	0.978	9.884	1.572
$\overline{\Delta}$, Full	0.893	7.087	0.624

good results in terms of both JI and UC. Similar results in terms of JI and UC are obtained using the proposed SGM function, while the SGM outperforms the JI in terms of excess capacity. The system which minimizes $\overline{\Delta}$ achieves relatively good results in terms of EC and UC, while performing poorly in terms of fairness. This is due to the fact that some of the beams, especially those with lower capacity request, are in some case left in starvation with no capacity allocated. Such results are in line with the expectations, since the $\overline{\Delta}$ system tries to minimize the capacity mismatch without taking fairness into account, while the JI system maximizes fairness disregarding the capacity mismatch. However, since the system is well dimensioned to meet the capacity request for the off-peak traffic, also the result in terms of UC for the JI system is relatively good. In the off-peak

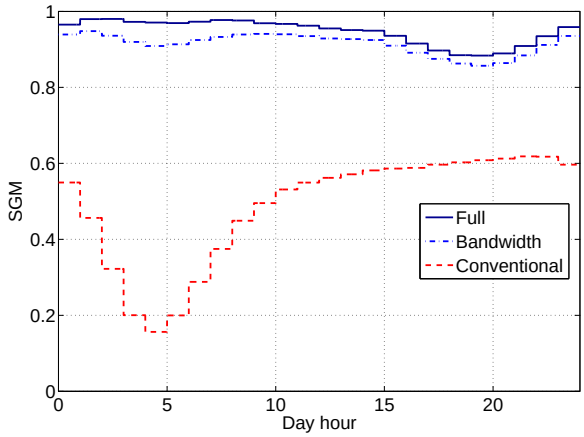

Fig. 10: SGM plotted against the hour of the day obtained applying the proposed algorithm to two payloads each having a different degree of flexibility (bandwdth and power, bandwidth only). The performance of a conventional (i.e., not optimized) payload is also shown.

hour the proposed SGM function achieves a result which is a compromise between the two systems and achieves relatively good results in terms of both fairness and capacity mismatch. When the traffic request increases, as shown in Table IV, the usefulness of the proposed SGM function becomes more evident. As a matter of facts, it achieves the same fairness as the JI system while having lower UC and lower EC. The $\overline{\Delta}$ system shows good results in terms of EC and UC, but has much worse results in terms of fairness.

B. Time Varying Capacity Request

In order to evaluate the performance of the proposed solution in case of time-varying requested capacity, we performed extensive simulations increasing the number of user terminals to 10^4 (50 terminals per beam) and using the requested capacity model presented in [6, section III-E]. The results of the simulations are shown in Fig. 10, where the SGM is plotted against the hour of the day. The results obtained applying the proposed algorithm to the two most promising payloads, i.e. the one with full flexibility and the one with only bandwidth flexibility, are shown. The performance of a conventional (i.e., not optimized) payload is also shown for comparison. The optimization algorithm is run once per hour so as to follow the traffic request variation. In the initial state of the optimization algorithm for the time 00:00 the bandwidth and power allocation of all payloads are set as in the conventional payload. The initial state for each subsequent hour has been set equal to the solution of the previous hour. This was done in order to save computational resources. In fact, since the traffic distribution across beams shows only limited changes within an hour, using the solution obtained in the previous hour as a starting point is likely to lead to a better results for a given amount of computational resources with respect to the case in which the conventional payload allocation is used as a starting point. From the figure it can be seen how the proposed algorithm achieves a significant gain in terms of SGM for both flexible payloads with respect to the conventional case. The advantage of having both power and bandwidth flexibility with respect to having bandwidth flexibility only appears to be relatively limited throughout the whole day. The largest gain, around 10%, can be observed at around 5:00, in correspondence to a period of low traffic request. Such gain is due mostly to the lower excess capacity of the fully flexible payload, which confirms what presented in Table I. The low SGM values achieved by the conventional payload in the same period are also due mostly to excess capacity. During periods of high traffic demand (around 19:00) the flexible payloads show a minimum in the SGM, which is caused by the high capacity request and possibly by an under-dimensioning of the system. Despite of this, a gain of around 40% with respect

to the conventional payload can be observed. The fact that the performance of the fully flexible payload are at least as good as those obtained by the payload with only bandwidth flexibility, both evaluated in the respective optimal configurations, is intuitive, since the Full payload can, at worst, reproduce the optimal allocation of the Bandwidth one. However, this is not guaranteed if non-optimal feasible points are chosen. The fact that, as shown in Fig. 10, the Full always upper-bounds the Bandwidth one is an indication of the goodness of the algorithm.

VI. Conclusions

We studied the problem of radio resource management in multibeam satellite systems under realistic conditions. A novel objective function for the optimization problem has been introduced with the aim to match the requested capacity across the beams as close as possible while taking fairness into account. After showing that, at least for some constraints imposed by the payload, the proposed function is not convex, we proposed a stochastic optimization algorithm to minimize such function. The algorithm is based on a modified version of the simulated annealing algorithm. A detailed complexity analysis of the proposed algorithm has been presented. The algorithm has been applied to three payloads having different degrees of flexibility, namely flexibility both in bandwidth and power, in bandwidth only and in power only. Realistic payload models, antenna pattern, co-channel interference and requested traffic distribution were used in the simulations. Our results show that the proposed approach is more efficient than the traditional conventional payload in matching the requested capacity across the beams and leads to interesting results both under low and high traffic demand. From the results it also emerges that power flexibility shows limited performance enhancement with respect to a conventional payload design, while bandwidth flexibility definitely leads to better results. If the two kinds of flexibility are combined, a limited improvement in performance with respect to the bandwidth-flexible case is observed. The goodness of the proposed optimization approach has been supported by measuring different figures of merit traditionally used in this kind of analysis, namely missing capacity, excess capacity and Jain index. Furthermore, we showed that adopting the proposed objective function leads to good results compared to other relevant objective functions (namely Jain Index and capacity mismatch) in low traffic conditions, while outperforms such functions when higher traffic requests are considered. Finally, we performed extensive simulations to test the proposed solution in the case of time-varying capacity request. The results show significant gains with respect to the conventional payload in the whole considered time span.

Appendix

Proof of Proposition 3.1

In order to prove that the SGM is in general not concave we look for a setup in which this is not verified. We recall that a function $g(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R}$, is concave if and only if its domain $\mathcal{D}$ is a convex set and $\forall \lambda \in [0,1], \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$ the following holds [26]:

$$\lambda g(\mathbf{x}_1) + (1-\lambda) g(\mathbf{x}_2) \leq g(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2).$$

Let us consider a system with 4 beams, two unitary-bandwidth bandwidth chunks and two TWTAs. Each TWTA is connected to two beams. Bandwidth and power can be allocated by a TWTA independently of the other TWTA. Assume that one user per beam is present. Now consider the following channel matrix:

$$\mathbf{H} = \frac{2}{3}\mathbf{I} + \frac{1}{3}\mathbf{1}^{4\times 4}$$

where $\mathbf{1}^{k\times l}$ is a $k \times l$ all-one matrix while $\mathbf{I}$ represents the identity matrix. Element $h_{i,j}$ of matrix $\mathbf{H}$ represents the channel coefficient from beam j to user in beam i. Element $h_{i,i}$ represents the gain of the reference beam while $h_{i,j}$, $i \neq j$ represents the channel gain of the interfering beam j as observed by the user in beam i. Finally, consider the following vector of requested capacities: $(T_r^1\ T_r^2\ T_r^3\ T_r^4) = (1\ 2\ 0.5\ 0)$. The spectral efficiency is evaluated using the Shannon capacity formula for the AWGN channel and calculating the SINR using expression (11) [11]. Consider now a family of feasible points characterized by the following bandwidth matrix:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Let us now consider two different feasible points defined by $\mathbf{B}$ and the following power allocation vectors (in linear scale) each containing the OBO of the two TWTA:

$$\mathbf{v}_1 = \begin{pmatrix} 0.2 \\ 0.1429 \end{pmatrix}$$

$$\mathbf{v}_2 = \begin{pmatrix} 0.3333 \\ 0.5 \end{pmatrix}.$$

Let $\text{SGM}(\mathbf{v}_1, \mathbf{B})$ and $\text{SGM}(\mathbf{v}_2, \mathbf{B})$ be the SGM values obtained with the two configurations. Let us now define $\lambda_1 = 0.3$ and $\lambda_2 = 1 - \lambda_1$. Finally we have:

$$\begin{aligned} \lambda_1 \text{SGM}(\mathbf{v}_1, \mathbf{B}) + \lambda_2 \text{SGM}(\mathbf{v}_2, \mathbf{B}) &= 0.5975 \\ &> 0.5964 = \text{SGM}(\lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2, \mathbf{B}) \end{aligned} \tag{22}$$

which concludes the proof.

ACKNOWLEDGMENT

The research activity reported in the present paper has been partly funded by the ESA/ESTEC ARTES 5.1 project "MultiBeam-RRM Efficient load-aware dynamic global radio resource management in multi-spotbeam broadband satellite networks" (contract # 4000109718/2014/NL/WE).

REFERENCES

[1] D. Serrano-Velarde, E. Lance, and H. Fenech, "Novel dimensioning method for high-throughput satellites: forward link," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 50, no. 3, pp. 2146–2163, July 2014.

[11] Note that the DVB-S2X standard has a high spectral efficiency which actually approaches Shannon capacity for a practically acceptable packet loss rate over a wide SNR range.

[2] H. Fenech, A. Tomatis, S. Amos, V. Soumpholphakdy, and J. S. Merino, "Eutelsat HTS systems," *Int. J. of Satellite Commun. and Networking*, vol. 34, pp. 503–521, 2016.

[3] European Telecommunications Standards Institute (ETSI), "Digital Video Broadcasting (DVB), user guidelines for the second generation system for broadcasting, interactive services, news gathering and other broadband satellite applications (DVB-S2)," ETSI TR 102 376 V1.1.1, Tech. Rep., Feb. 2005.

[4] ——, "Digital Video Broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applicaitons. (DVB-S2X)," ETSI EN 302 307 part 2, Tech. Rep., Oct. 2014.

[5] M. A. V. Castro and G. S. Granados, "Cross-layer packet scheduler design of a multibeam broadband satellite system with adaptive coding and modulation," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 1, pp. 248–258, Jan 2007.

[6] X. Alberti, J. Cebrian, A. D. Bianco, and N. Alagha, "System capacity optimization in time and frequency for multibeam multi-media satellite systems," in *Advanced Satellite Multimedia Systems Conf.*, Cagliari, Italy, Sep. 2010, pp. 226–233.

[7] J. Lizarraga, P. Angeletti, N. Alagha, and M. Aloisio, "Flexibility performance in advanced Ka-band multibeam satellites," in *IEEE Int. Vacuum Electronics Conf.*, Apr. 2014, pp. 45–46.

[8] J. Lei and M. A. Vazquez-Castro, "Joint power and carrier allocation for the multibeam satellite downlink with individual SINR constraints," in *IEEE Int. Conf. on Commun.*, Cape Town, South Africa, May 2010, pp. 1–5.

[9] J. Lei and M. Vzquez-Castro, "Multibeam satellite frequency/time duality study and capacity optimization," *J. of Commun. and Networks*, vol. 13, no. 5, pp. 472–480, Oct 2011.

[10] J. P. Choi and V. W. S. Chan, "Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks," *IEEE Trans. on Wireless Commun.*, vol. 4, no. 6, pp. 2983–2993, Nov. 2005.

[11] ——, "Resource management for advanced transmission antenna satellites," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 3, pp. 1308–1321, Mar. 2009.

[12] T. Qi and Y. Wang, "Energy-efficient power allocation over multibeam satellite downlinks with imperfect CSI," in *Int. Conf. on Wireless Commun. Signal Processing*, Nanjing, China, Oct. 2015, pp. 1–5.

[13] A. Aravanis, B. Shankar, M.R., P. Arapoglou, G. Danoy, P. Cottis, and B. Ottersten, "Power allocation in multibeam satellite systems: A two-stage multi-objective optimization," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 6, pp. 3171–3182, June 2015.

[14] B. G. Evans, P. T. Thompson, and A. Kyrgiazos, "Irregular beam sizes and non-uniform bandwidth allocation in HTS," in *AIAA Int. Commun. Satellite Systems Conf.*, Florence, Italy, Sep. 2013, pp. 1–7.

[15] S. Salcedo-Sanz, R. Santiago-Mozos, and C. Bousoño-Calzón, "A hybrid Hopfield network-simulated annealing approach for frequency assignment in satellite communications systems," *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 2, pp. 1108–1116, Apr. 2004.

[16] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[17] P. Ferreira, R. Paffenroth, A. Wyglinski, T. H. S. Biln, R. Reinhart, and D. Mortensen, "Multi-objective reinforcement learning for cognitive radio–based satellite communications," in *AIAA Int. Commun. Satellite Systems Conf.*, Cleveland, OH, Oct. 2016.

[18] T. Olwal, K. Djouani, and A. Kurien, "A survey of resource management towards 5G radio access networks," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 3, pp. 1656–1686, 2016.

[19] H. Zhang, C. Jiang, N. Beaulieu, X. Chu, X. Wang, and T. Quek, "Resource allocation for cognitive small cell networks: A cooperative bargaining game theoretic approach," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 6, pp. 3481–3493, June 2015.

[20] G. Cocco, T. de Cola, M. Angelone, and Z. Katona, "Radio resource management strategies for DVB-S2 systems operated with flexible satellite payloads," in *Advanced Satellite Multimedia Systems Conf.*, Palma de Mallorca, Spain, Sep. 2016.

[21] G. Maral and M. Bousquet, *Satellite Communications Systems: Systems, Techniques and Technology*, 5th ed. Chichester, England: John Wiley and Sons Ltd, 2009.

[22] A. Kalai and S. Vempala, "Simulated annealing for convex optimization," *Mathematics of Operations Research*, vol. 31, no. 2, pp. 253–266, 2006. [Online]. Available: http://dx.doi.org/10.1287/moor.1060.0194

[23] N. Koblitz, *A Course in Number Theory and Cryptography*, 2nd ed., S. Axler, F. Gehring, and K. Ribet, Eds. Springer, 1998.

[24] M. Aloisio, P. Angeletti, E. Casini, E. Colzi, S. D'Addio, and R. Oliva-Balague, "Accurate characterization of twta distortion in multicarrier operation by means of a correlation-based method," *IEEE Trans. on Electron Devices*, vol. 56, no. 5, pp. 951–958, May 2009.

[25] http://www.ticra.com/products/software/grasp.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.