QoE-Aware Cost-Minimizing Capacity Renting for Satellite-as-a-Service enabled Multiple-Beam SatCom Systems

Teweldebrhan Mezgebo Kebedew, *Student Member, IEEE,* Vu Nguyen Ha, *Member, IEEE,*
Eva Lagunas, *Senior Member, IEEE,* Joel Grotz, *Senior Member, IEEE,* Symeon Chatzinotas, *Fellow, IEEE*

***Abstract*—The advent of Satellite as a Service (SaaS) platforms has empowered satellite service providers (SPs) to rent portions of satellite capacity from infrastructure providers (IPs) to cater to the diverse demands of their users across multiple satellite services. To effectively manage costs and maintain a high Quality of Experience (QoE) for numerous concurrent connections, SPs should secure flexible capacity from IPs. However, the irregular and unpredictable nature of traffic demands from various applications complicates the capacity-renting framework. This study presents a dynamic capacity allocation framework that efficiently handles diverse traffic flows with varying arrival rates, aiming to minimize rental costs while meeting blocking probability and QoE requirements. Utilizing the $M_t/M_t/1$ queuing model and a continuous-time Markov chain, the technical designs are framed as a statistical optimization problem. In this context, the system waiting-queue lengths are estimated using the transient probabilities of Kolmogorov equations. Subsequently, cumulative distribution functions are employed to re-formulate this stochastic optimization problem into a convex form, which can be tackled through the Lagrangian duality method.Through extensive simulations and numerical assessments, we illustrate our method's efficacy, with the proposed algorithm outperforming benchmarks by reducing costs by up to 9.85% and 3.1%.**

***Index Terms*—Time-varying queuing, capacity allocation, blocking probability, QoE-based optimization, satellite as a service.**

I. Introduction

SATELLITE communication (SatCom) systems are becoming important to provide global connectivity with a wide range of applications requiring high availability and resilience to critical areas that are unreachable by the current terrestrial networks [2]–[4]. The growing attention on the dynamic SatCom systems supporting service-aware communications and seamless coverage, as well as its associated implementation costs, has inspired a number of big names in the industry to develop new "satellite-as-a-services" (SaaS) business platforms [5], [6]. Such a platform comprises infrastructure providers (IPs) - the owners of the SatCom systems, and the service providers (SPs) which may rent certain satellite capacity from the IPs to provide different services to their end users [1], [5], [7], [8]. This novel paradigm can enable the IPs to access new markets and it can release the SPs from investing a huge amount of money in order to build dedicated satellites [9]. With the fluctuating demands from various applications and services, there's a growing interest in developing a flexible capacity renting mechanism for the SPs. This mechanism is expected to allow the SPs to obtain capacity as needed rather than committing to a fixed amount. Such flexibility can boost network utilization, enhance revenue for the IPs, and help the SPs save the costs while ensuring the user Quality of Experience (QoE) [10]. However, to establish such an optimal capacity renting framework between the IPs and SPs, one should carefully manage and analyze a vast number of simultaneous data flows, each with distinct QoE requirements [11]. Specifically, the SPs must swiftly make provisions on the required capacity while still guaranteeing the desired QoE for their users. On the SatCom IPs side, to realize the dynamic capacity allocation mechanism, the advanced digital transparent payloads (DTP) using *"defacto"* platform can be deployed [12], [13]. Based on this, the satellites can assign capacity across beams effectively and route traffic of various types efficiently.

To ensure effective network resources renting from the IPs and optimal capacity allocation for customer communication services, the SPs have to cope with multiple challenges in maximizing their profit [9], [14]. Herein, the business problems at the SPs include (but are not limited to) capacity estimation accommodating the irregular and unpredictable *"time-varying"* data traffic from heterogeneous apps/services, and efficient management of rented resources during both peak and off-peak periods to reduce costs [15], [16]. Additionally, maintaining or expanding the customer base is a significant business challenge for both the IPs and SPs, and satisfying the diverse QoE requirements of customers is a key point of resource management that should be addressed carefully. Furthermore, the IPs and SPs should also establish an agreement detailing the speed of capacity adjustments and the maximum capacity SPs are allowed to use [7], [14]. Consequently, the development of a dynamic, cost-effective, QoE-aware network resource renting mechanism in SatCom systems has emerged as an interesting and compelling research topic for both IPs and SPs. Facilitated by the challenging problem, this paper focuses on designing a new QoE-aware cost-minimizing capacity-renting framework in SaaS-enabled multi-beam SatCom systems.

In the realm of wireless communication, the QoE at users can

The preliminary results of this manuscript were presented in VTC-Fall 2022 [1]. This work has been supported by the Luxembourg National Research Fund (FNR) partially under the project INSTRUCT (IPBG19/14016225/INSTRUCT) and partially under the project SENTRY (C23/IS/18073708).

T. M. Kebedew, V. N. Ha, E. Lagunas, and S. Chatzinotas are with the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, 1855 Luxembourg Ville, Luxembourg. (e-mail: teweldebrhan.kebedew@uni.lu; vu-nguyen.ha@uni.lu; eva.lagunas@uni.lu; Symeon.Chatzinotas@uni.lu).

J. Grotz is with SES, Chateau de Betzdorf, Betzdorf 6815, Luxembourg (e-mail: Joel.Grotz@ses.com).

be affected by a variety of factors, ranging from signal strength, data rate consistency, and connection reliability, to service interruptions [17]. Among these aspects, a critical determinant of QoE in the SatCom schemes is the end-to-end latency where extended delay can significantly compromise a user's overall experience. While all QoE issues are essential, reducing the latent period is paramount [17]. It is worth noting that the end-to-end data transmission latency encompasses both the signal propagation delay and the waiting time during which data resides in the system buffer before transmission [18], [19]. However, in SatCom schemes, mitigating the propagation delay poses a formidable challenge due to the Line-of-Sight (LoS) connection and almost-fixed distances between the satellite and user terminals. This inherent physical limitation demands innovative solutions to ensure consistent and efficient communication, enhancing the overall user experience.

In light of this challenge, our study focuses on addressing the QoE issue by modeling it in terms of the waiting time of data packets in the system buffer. This approach underscores the significance of minimizing delays to improve overall user satisfaction. At a specific time instance, the waiting time of one beam is related to the stochastic queuing length which can be managed by dynamically allocating the capacity for the data transmission corresponding to this beam when the data arrival rates vary. In general, a longer queuing length returns a higher network congestion probability and a longer service time that the users may suffer because the data spends more time in the buffer before being processed and transmitted. When the queuing length approaches a limit, the waiting time may exceed users' tolerance; hence, the QoE rate can degrade. In addition, when the queue length violates the maximum buffer size of the system in some critical scenarios, the satellite system can be overloaded and the operating system can be blocked, resulting in packet loss. To cope with such an issue and conserve the required QoE, one demands a higher allocated capacity to increase the service rate. On another hand, letting the system operate with very short queuing lengths may imply that an over-needed amount of capacity is allocated, and the renting cost must increase. Then, our work aims to propose a novel dynamic capacity planning model that minimizes the total renting cost of SPs while maintaining a target blocking probability of the system and a target queuing delay requirement of customers.

A. Related Works

The majority of existing frameworks for SatCom resource allocation have primarily focused on maximizing overall power and spectrum utilization efficiency. Different techniques have been used, such as non-convex optimization for flexible power and capacity assignment [20], beam illumination and selective precoding [21], and joint beam selection and precoding [22]. The work in [20] focused on satellite–user association-oriented capacity allocation to minimize the total uplink transmit power for integrated satellite-terrestrial networks (ISTN). Research on QoE-aware dynamic capacity allocation to maximize user satisfaction in Orthogonal Frequency Division Multiple Access (OFDMA) terrestrial networks considering time-varying channels has been conducted [23], [24]. These approaches aim to satisfy the overall demand by improving power, capacity, or both utilization efficiency. However, all previous works consider average beam demand which does not change over time. In addition, the cost of satisfying this demand, the QoE of users, and the amount of system capacity that remains unused is not well-documented in the literature. Other studies have discussed profit opportunities associated with 5G infrastructure dynamic leasing [14] and revenue management in SatCom systems [25].

Previous research has also applied different queue models for end-to-end latency estimation, capacity allocation, packet loss minimization, and buffer bloat prevention in various wireless communication systems. For example, the $M/G/1$ queue model has been used to estimate transient queue length [26], latency estimation [27]–[29], the $D/D/K$ queue model for dynamic buffer sizing [30], the $G/G/1$ queuing model for Quality of Service (QoS) analysis [31], and the $M/G/\infty$ queue model to estimate the minimum required system capacity [32]. While these studies use different queue models to maintain QoS in different wireless systems, they fail to consider the QoE of users and the time-varying nature of average arrival rates.

In other areas, time-varying queue model-based resource allocation has been explored [33]. For example, in [34], [35] the application of different time-varying queue models in large-scale service systems such as customer contact centers and hospital emergency departments is discussed. In these works, iterative staffing algorithms (ISA) are developed to optimize the staff levels at the customer center to satisfy the stochastic waiting time requirement of customers. Herein, the state-of-the-art $M_t/M/S(t)$ queue model is exploited where the M_t indicates time-varying Poisson arrivals, and the $S(t)$ indicates time-varying servers and the M indicates a constant serving capacity of the servers in every staffing interval. Then, an efficient time-varying human-resource-management framework is established by iteratively determining the staff level at a specific time. Similarly, the authors in [36] also employed $M_t/M/S(t)$ to develop a *Deep-Reinforcement-Learning* based framework to optimize the beam-hopping strategy adapting with time-varying data traffic flows coming to the multiple-beam SatCom systems. Next, [37] describes how time-varying $M_t/G/\infty$ queue models can be applied to staffing and capacity planning of cloud services and protective equipment management in hospitals during the outbreak of a disease. All works given in [34], [35], [37] focus on a fixed processing rate at one serve (staff). On a different approach, the changeable processing rate mechanism is studied in [38] where the $M_t/M_t/1$ queue model is used for efficient resource allocation in industrial production by estimating the stochastic probability of the system queue length. However, this work mainly focuses on developing the admission control strategy for network-slicing systems. Consequently, there is a very limited number of works considering multiple traffic flows accessing satellite systems as time-varying queue models and exploiting this to develop a QoE-aware dynamic capacity renting and allocation mechanism for the SPs to in multiple-beam SatCom systems. Therefore, the work in this paper aims to fill this gap in the literature.

B. Contributions

Our paper aims to propose a novel QoE-aware flexible capacity renting framework for SaaS-enabled multiple-beam

SatCom systems to effectively manage the renting costs at the SPs. By employing the stochastic queuing theory, we first formulate the problem as a stochastic optimization problem. This allows us to examine the impact of system and user requirements on the rent cost and the trade-off between them, thereby assisting SPs in predicting the capacity they need to request IPs to maintain a satisfied customer base while minimizing rent costs. Our key technical contributions in this article can be summarized as follows:

- First, we express the traffic flows using a time-varying $(M_t/M_t/1)$ queuing model and estimate the stochastic queue length of packets waiting in the system by using the continuous-time Markov chain (CTMC). The analysis results are then employed to formulate a stochastic optimization problem for QoE-aware dynamic capacity allocation that includes queue status-based dynamic spectrum sharing among adjacent beams of the same cluster to minimize cost. The problem aims to help SPs to be able to efficiently allocate capacity and reduce un-utilized capacity as well as their rent costs.
- Next, we estimate the stochastic blocking probability and the probability of violating the waiting-time requirement over the observation period, which allows SPs to predict the impact of capacity allocation decisions on user experience. We further analyze the trade-off between maintaining user satisfaction and capacity rent cost in SatCom systems will assist SPs in determining the optimal balance between these two objectives, based on which we provide a closed-form solution based on Lagrangian duality making use of the estimated blocking probability.
- For comparison purposes, we introduce a greedy algorithm and modify the ISA frameworks in [34], [35] to suit our design requirements. The proposed algorithms are validated through numerical results and Monte Carlo simulations using practical simulation parameters. The numerical and simulation outcomes have effectively confirmed the theoretical soundness of our proposed frameworks.

In summary, our proposed approach is based on the stochastic queuing theory and aims to assist satellite service providers in predicting the capacity they need to maintain a satisfied customer base while minimizing rent costs by using a QoE-aware dynamic capacity allocation model. Preliminary studies related to this objective were presented in [1]. This current work further extends our previous results by regarding multi-beam settings with cross-beam interference avoidance constraints, presenting robust theorem analysis, and delivering more solid numerical and simulation demonstrations alongside added benchmark comparisons. The rest of the paper is organized as follows. In Section II, the system model and problem formulation are described. The queuing stochastic analysis and problem approximation are discussed in Section III. In Section V, the numerical results are discussed. Finally, Section VI concludes the paper. For notation, scalars are represented in a normal *italic* font such as W and u, while the vectors and matrices are in **bold**, i.e., $\mathbf{W}$ and $\mathbf{u}$. In addition, sans serif font is utilized to denote the "suffix" abbreviations, such as, W^{total}, Q_{max}. For ease of reference, a list of key notations used in this paper is provided in Table I.

TABLE I: List of Key Notations

Notation	Definition
B	Total number of virtual beams
C	Number of adjacent beam clusters
K	Number of cycles
L	Normalized packet length (bits) (bps)
M	Number of time slots per cycle K
$P_{n,b}(t)$	Probability of n packets in beam b at time t
$\hat{P}_{n,b}(t)$	Approximated number of packets in beam b at time t
$\bar{P}_{\mathsf{Blk}}$	Target blocking probability
$\bar{P}_{\mathsf{QoE}}$	Target QoE requirement violation probability
Q_{max}	Maximum buffer length
Q_{QoE}	Target queue length
$Q^b(t)$	Queuing length of data flows in beam b at time t
S	Number of data flows per beam
S_k^b	Number of capacity packages in beam B at cycle K
T	Total observation time (seconds)
T_{TS}	Time slot duration (seconds)
$\mathbf{U}$	Adjacency matrix
$\mathbf{u}^c$	c-th row of $\mathbf{U}$
$V^b_{\mathsf{QoE},k}$	Max QoE violating probability for beam b at cycle K
$V^b_{\mathsf{bl},k}$	Maximum blocking probability in beam B at cycle K
$\mathbf{W}$	Matrix containing all W_k^b's
W^{total}	Total available capacity of the satellite (bps)
$W(t)$	Total rented capacity at time t
$W^b(t)$	Capacity rented for beam b at time t (bps)
W_k^b	Capacity rented for beam b at cycle k (bps)
$\mathbf{W}_k$	Vector of rented capacity at cycle k (bps)
$\lambda_s^b(t)$	Arrival rate data packets of flow s to beam b
β^b, ζ^c	Lagrangian multipliers
$\rho^b(t)$	Beam b utilization at time t
$\Lambda^b(t)$	Total arrival rate to beam b
γ	Price per Mbps

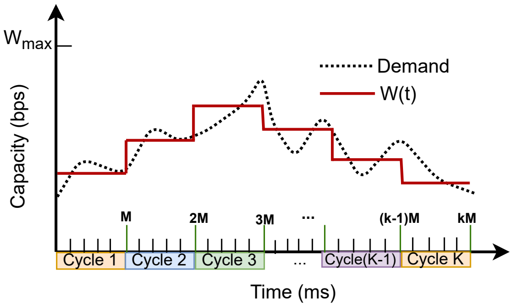


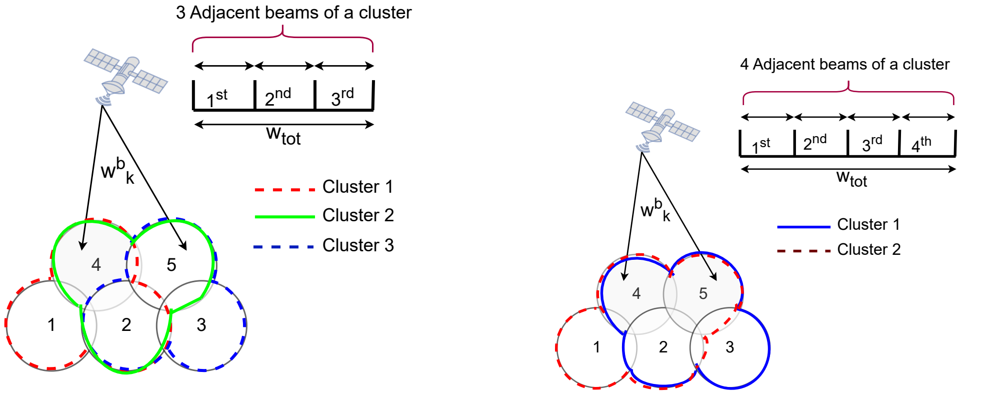
Fig. 1: Capacity allocation for time-varying demand.

II. System Model and Problem Formulation

We examine a SaaS platform in a multi-beam geostationary satellite (GEO) communication system where the overall available capacity of a GEO satellite is owned by an IP that can be rented by SPs[1] The study focuses on a scenario where a specific SP[2] rents a time-varying amount of capacity from the IP to provide broadband services to multiple users randomly distributed across B beams. Let $W(t)$ (bps) represent the total capacity rented by the SP from IP to serve all traffic flows across all beams. Each beam is assigned a portion of this capacity, denoted by $W^b(t)$, which can range from 0 to $W(t)$, i.e.

[1] It is worth noting that the results can be applied to other SaaS-enabled Non-geostationary satellite (NGSO) systems with multiple satellites as well.

[2] The work can be efficiently scaled up for a general scheme consisting of multiple SPs as well.



(a) Considering 3 adjacent beams per cluster.

(b) Considering 4 adjacent beams per cluster.

Fig. 2: The footprints of clusters containing J adjacent beams with non-overlapping frequency allocation.

$0 \leq W^b(t) \leq W(t)$. The network operates in a time-slot manner, with each time slot having a duration of T_{TS} as the transmission time. Due to possible processing speed capability limitations of DTP reconfiguration combined with the consequent signaling through the tracking, telemetry, and command system (TT&C) links [12], it is assumed that $W^b(t)$ remains constant for a cycle duration of M time slots. $W^b(t)$ can only be reset at time-slot indices $t_c \in \{0, M, 2M, ..., kM, ...\}$ and $W^b(t) = W^b(kM)$ if $t \in ((k-1)MT_{\mathrm{TS}}, kMT_{\mathrm{TS}}]$ where $k = 1, 2, ..., K$. We refer to $W^b(kM)$ as W^b_k, which represents the allocated capacity of beam b in cycle k. The design framework considers a monitoring period of K cycles, equivalent to a total operation time of $T = KMT_{\mathrm{TS}}$ (seconds).

Let $f(X)$ stand for the cost function corresponding to X (bps) that the SP rents from IP. Normally, $f(X)$ is a monotonic increasing function with respect to X which reflects the idea that renting more capacity incurs more rental cost. Then, the total renting cost that the SPs have to pay to the IP during T seconds can be expressed as,

$$F_{SP} = f\left(\int_0^T W(t)dt\right) = f\left(\sum_{b=0}^{B}\sum_{k=0}^{K-1} MT_{\mathrm{TS}} W^b_k\right). \quad (1)$$

A. Capacity Allocation and Frequency Reuse

To address the strong cross-beam interference between adjacent beams, a dynamic multi-color capacity allocation policy is employed in this multi-beam transmission system. This differs from traditional color-reuse schemes, where the spectrum is equally distributed. Here, the spectrum is freely allocated to beams, ensuring that different and non-overlapping frequency bands are assigned to two arbitrary adjacent beams. It follows that the sum of capacity assigned to any cluster of J adjacent beams must not exceed the maximum available spectrum band capacity. For instance, Figs. 2a and 2b demonstrate settings of 3 and 4 adjacent-beam clusters with non-overlapping spectrum.

For a specific beam pattern, let C be the number of available J-adjacent-beam clusters and $\mathbf{U} \in \{0,1\}^{C\times B}$ be the adjacency matrix. In the c^{th} row of $\mathbf{U}$, only J elements corresponding to the indices of J adjacent beams of cluster c are set to one, while the others are set to zero. The dynamic multi-color reuse capacity allocation requirement can then be described by the following constraint:

$$\mathbf{U}\mathbf{W}_k \leq W^{\mathrm{total}} 1_{C\times 1}, \quad (2)$$

where $\mathbf{W}_k = \left[W^1_k, ..., W^b_k, ..., W^B_k\right]^T$, W^{total} (bps) indicates the maximum reusable capacity available per cluster, and $1_{C\times 1}$ stands for a one-vector with size of $C \times 1$.

B. Queuing Model

From user perspective, a single device can generate multiple data packets that correspond to various applications operating on it concurrently. It's assumed that packets corresponding to a specific application type have identical packet sizes. Consequently, the number of packets resulting from a particular application, which is run by several devices concurrently, can be consolidated into a traffic flow with a projected arrival rate. This number of packets adheres to a random process, with the estimated arrival rate serving as the mean value [39]–[41]. This work deals with heterogeneous time-varying traffic rates generated by different applications such as voice, video streams, and web browsing. To handle this, the demands from end users are modeled as multiple queues accessing each beam. The data flows are classified based on their corresponding statistical parameters, such as arrival rates and packet lengths. In this model, the arrived packets are processed based on the basic first-come first-served strategy.

Consider S data-flows corresponding to S services tending to access each of B beams as shown in Figure 3. We further assume that the flow s carrying data packets of L_s-bit length comes to beam b at a time t following an independent Poison process[3] [39]–[42] with a time-varying arrival rate of $\lambda_s(t)$, i.e., $\lambda_s(t)$ is the number of packets that changes over the time. The total arrival rate in bits per second to a beam becomes

$$\Lambda^b(t) = \sum_{s=1}^{S} \lambda^b_s(t) L_s. \quad (3)$$

[3] The traffic-flow arrival rates of some typical use cases for the next-generation communication services, such as Internet of Things with small packets and virtual-image communication, have been reported to follow the Poisson process [39]–[42].

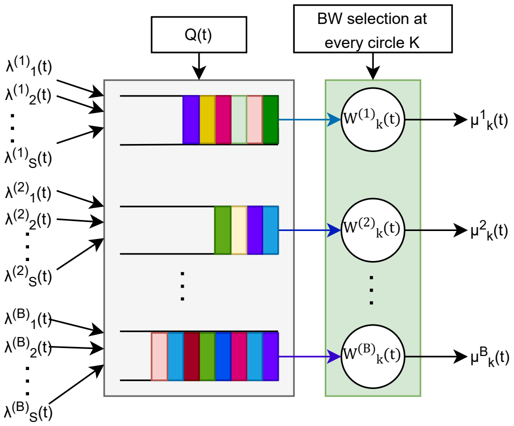


Fig. 3: Capacity allocation for a queued flow of packets in multiple beam satellite networks.

Suppose that a packet of L bits is a normalized processing unit in this SatCom system. As a result, the arrived data can be divided into packets of L bits for transmission. Consequently, the total arrival rate in terms of the number of packets at beam b can be determined as

$$\lambda^b(t) = \Lambda^b(t)/L. \tag{4}$$

Additionally, the service rate for all flows entering beam b at any time t, which is estimated as the number of packets based on the corresponding allocated capacity, can be expressed as,

$$\mu^b(t) = W^b(t)T_{\mathsf{TS}}/L. \tag{5}$$

Remark 1. *The users' demands in terms of QoS can be captured by the traffic flow model. For instance, a higher transmission rate can be equated to an elevated arrival rate ($\lambda_s^b(t)$ is scaled-up) or an extended packet size (L_s is set to a larger value). Moreover, rapid time-varying requirements can be depicted by a more fluctuating function of the time-varying arrival rate setting, $\lambda_s^b(t)$.*

C. QoE Requirement and Problem Formulation

In this section, we aim to ensure that the allocated capacity for every beam does not become a bottleneck violating the required QoE. Here, we probabilistically model the QoE requirements in the manner of transmission delay (waiting time) for the users in each beam with the queuing length of packets available in the data buffer corresponding to that beam transmission.

Specifically, let $Q^b(t)$ denote the queuing length of data packets stored in the buffer of beam b at time t. Then, one denotes $P_{n,b}(t) = Pr(Q^b(t) = n)$ as the probability that there are n packets in the buffer of beam b at time t. As discussed in [33], due to the time-varying arrival rate models, $P_{n,b}(t)$ can be expressed as a function of t. Assuming that users in all beams have the same waiting time tolerance which corresponds to a QoE threshold of queuing length Q_{QoE}. The experience of network utilization is considered "acceptable" by the users if the probability of that such threshold is violated is less than a commitment factor $\bar{P}_{\mathsf{QoE}}$, i.e., $0 < \bar{P}_{\mathsf{QoE}} < 1$. Therefore, the design in this work focuses on keeping the probability that the queue length surpasses Q_{QoE} packets over the window time of $[0, T]$ less than $\bar{P}_{\mathsf{QoE}}$ for all the beams, which can be expressed as

$$\frac{1}{T}\int_0^T Pr\{Q^b(t) \geq Q_{\mathsf{QoE}}\}dt \leq \bar{P}_{\mathsf{QoE}}, \forall b. \tag{6}$$

One further assumes that the length of every beam buffer is limited by Q_{max}, so-called the maximum buffer length. Herein, it also needs to ensure that the queue length at every beam does not surpass Q_{max} beams as much as possible, otherwise, the processing of all data flows to the particular beam will be blocked. Regarding the network admission requirements, our design needs to maintain the blocking probability below a predetermined threshold for every time slot [42]. This requirement can be cast by the following constraint,

$$Pr\{Q^b(t) \geq Q_{\mathsf{max}}\} \leq \bar{P}_{\mathsf{Blk}} \quad \forall(t, b), \tag{7}$$

where $\bar{P}_{\mathsf{Blk}}$ is the target blocking probability. Taking into account that $Pr\{Q^b(t) \geq N\} = 1 - \sum_{n=1}^{N} P_{n,b}(t)$, our technical designs can be formulated into a statistical optimization problem as follows.

$$\min_{\mathbf{W}} \quad f\left(\sum_{\forall b}\sum_{\forall k} MT_{\mathsf{TS}}W_k^b\right) \tag{8a}$$

$$\text{s.t.} \quad \text{constraint (2)},$$

$$\sum_{n=0}^{Q_{\mathsf{max}}} P_{n,b}(t) \geq 1 - \bar{P}_{\mathsf{Blk}}, \forall(t, b), \tag{8b}$$

$$\frac{1}{T}\int_0^T \left(\sum_{n=0}^{Q_{\mathsf{QoE}}} P_{n,b}(t)\right) dt \geq 1 - \bar{P}_{\mathsf{QoE}}, \forall b, \tag{8c}$$

where $\mathbf{W}$ represents the matrix containing all W_k^b's. As observed, this presents a stochastic optimization problem wherein the constraints correspond to a random process. The primary challenge in resolving this problem stems from the statistical formulas articulated in constraints (8b) and (8c). In this context, while the problem data remains uncertain, the queuing model incorporating the Poisson process, as discussed in Section II-B, serves as the foundation for our solution framework.

Remark 2. *It is worth noting that $f(X)$ is an increasing function so the SP has to pay more if it rents more capacity. Hence, problem* (8) *is equivalent to the following,*

$$\min_{\mathbf{W}} MT_{\mathsf{TS}}\sum_{\forall b}\sum_{\forall k} W_k^b \qquad \textit{s.t. constraints } (2), (8b), \textit{ and } (8c). \tag{9}$$

III. Queuing Stochastic Analysis and Problem Approximation

A. Time-Varying Queuing Stochastic Brief Discussion

The total demand in each beam, which is the sum of arrivals of all data flows, varies over time and can be modeled as a continuously varying arrival rate. Consequently, the number of packets in the buffers in each time slot is a stochastic process. To model this behavior, we represent the number of packets at each time slot as a CTMC, as shown in Fig. 4. This stochastic

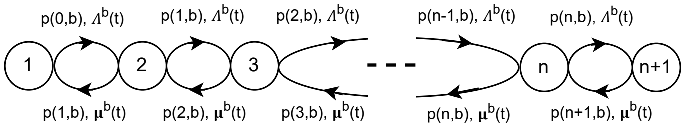


Fig. 4: Queue state transition diagram of beam b.

process can be described using a system of ordinary differential equations (ODEs) known as the Kolmogorov equations [33], and the queue length follows a birth-death process. The transient solution of the Kolmogorov equation yields the stochastic queue length values as:

$$\frac{\partial P_{0,b}(t)}{\partial t} = \begin{cases} -\lambda^b(t)P_{0,b}(t) + \mu^b(t)P_{1,b}(t), & \text{if } n = 0, \\ \lambda^b(t)P_{n-1,b}(t) + \mu^b(t)P_{n+1,b}(t) & \\ \qquad -(\lambda^b(t) + \mu^b(t))P_{n,b}(t), & \text{if } n > 0, \end{cases} \tag{10}$$

where $\frac{\partial P_{0,b}(t)}{\partial t}$ indicates the derivative of $P_{n,b}(t)$ at time t. However, since the Kolmogorov equation does not yield explicit solutions for the transition probabilities, various more suitable methods have been developed to approximate the solutions, as described in [33], [38], [43]. Denote $Q(t)$ as the queue length at time t. Then, the transient probabilities can be approximated by a cumulative distribution function given by:

$$F(Q, t) = Pr(Q(t) \leq Q). \tag{11}$$

The expected value of the queue distribution at the buffer of beam b provides,

$$\int_n^{n+1} F(Q, t)dQ \approx P_{n,b}(t). \tag{12}$$

The beam utilization at any time slot t is calculated as,

$$\rho^b(t) = \frac{\lambda^b(t)}{\mu^b(t)}. \tag{13}$$

The continuous-model approximation is exact in steady-state [38]. Therefore, the approximated probability of the availability of n packets in the buffer system of beam b at any time slot t can be expressed as

$$\hat{P}_{n,b}(t) = \int_n^{n+1} F(Q)dQ = \rho^b(t)^n(1 - \rho^b(t)). \tag{14}$$

B. Problem Approximation

This section exploits the continuous-model approximation in (14) to express the stochastic queue length in terms of capacity and express the constraints as a function of capacity W_k^b. Expressing $\mu^b(t)$ in (5) in terms of W_k^b for $t \in ((k-1)M, kM)$ and $k \in \{1, .., K\}$, one can rewrite (14) as,

$$\begin{aligned} \hat{P}_{n,b}(t) &= g_{n,b}(W_k^b, t) \\ &= \left(\frac{\sum_{s=1}^S \lambda_s^b(t)L_s}{T_{\mathsf{TS}}W_k^b}\right)^n \left(1 - \frac{\sum_{s=1}^S \lambda_s^b(t)L_s}{T_{\mathsf{TS}}W_k^b}\right), \quad \forall(k, b). \end{aligned} \tag{15}$$

Since the beams can not serve beyond their maximum capacity, the system utilization constraint needs to satisfy $\rho^b(t) \leq 1$ [38]. Otherwise, $\rho^b(t) > 1$ indicates the system is over-congested and users are blocked from accessing the corresponding beam. Letting Ω_k denote the set $\big((k-1)M, kM\big]$, one can restate problem (9) as,

$$\min_{\mathbf{W}} \quad MT_{\mathsf{TS}} \sum_{b=1}^B \sum_{k=1}^K W_k^b \tag{16a}$$

$$\text{s.t.} \quad \text{constraint (2)},$$

$$\sum_{n=0}^{Q_{\mathsf{max}}} g_{n,b}(W_k^b, t) \geq 1 - \bar{P}_{\mathsf{Blk}}, \forall k, \forall b \text{ and } t \in \Omega_k, \tag{16b}$$

$$\frac{1}{T}\sum_{k=1}^K \int_{(k-1)M}^{kM} \sum_{n=0}^{Q_{\mathsf{QoE}}} g_{n,b}(W_k^b, t)dt \geq 1 - \bar{P}_{\mathsf{QoE}}, \forall b, \tag{16c}$$

$$0 \leq \rho^b(t) \leq 1, \ \forall t, \forall b. \tag{16d}$$

IV. Dynamic Capacity Allocation Design

In this section, we aim to explain the step-by-step approach to find the optimal dynamic capacity allocated across all beams.

A. Problem Convexity Characterization

In order to solve problem (16), we first characterize its convexity. To begin with, we first define the lower bound capacity amount required per cycle in which $\mu^b(t)$ is fixed by considering the following proposition.

Proposition 1. *Constraints* (16b) *and* (16d) *in problem* (16) *can be merged into one constraint as*

$$W_k^b \geq \alpha_k^b = \max\left(\alpha_{k,1}^b, \alpha_{k,2}^b\right), \forall k, \forall b, \tag{17}$$

where $\alpha_{k,1}^b = \max_{t\in\Omega_k} L\lambda^b(t)/T_{\mathsf{TS}}$, $\alpha_{k,2} = \max_{t\in\Omega_k} Lg_{Q_{\mathsf{max}}}^{-1}(1 - \bar{P}_{\mathsf{Blk}}, t)/T_{\mathsf{TS}}$, *and* $g_{Q_{\mathsf{max}}}^{-1}(\bar{P}_{\mathsf{Blk}}, t)$ *is the inverse function of* $\sum_{n=0}^{Q_{\mathsf{max}}} g_{n,b}(W_k^b, t)$.

Proof: The proof is given in Appendix A ■

In the next move, based on the result of this proposition and the fact that (17) is a linear constraint, we state the convexity of problem (16) in the following theorem.

Theorem 1. *Problem* (16) *can be transformed into the following optimization problem which is convex,*

$$\min_{\mathbf{W^b}} \quad \sum_{\forall b} \sum_{\forall k} MT_{\mathsf{TS}}W_k^b \tag{18a}$$

$$\textit{s.t.} \quad \textit{constraint (17)},$$

$$\sum_{\forall k} z_k(W_k^b) \geq 1 - \bar{P}_{\mathsf{QoE}}, \tag{18b}$$

where $z_k(x) = MT_{\mathsf{TS}}/T - A_k^b/\big(Tx^{Q_{\mathsf{QoE}}+1}\big)$.

Proof: The proof is given in Appendix B ■

Thanks to Proposition 1 and Theorem 1, one can state that (16) is equivalent to convex problem (18). In the following section, a dynamic resource allocation algorithm is proposed by developing an optimization-based approach to obtain the optimal solution of this problem.

B. Duality-based Dynamic Capacity Allocation Algorithm

1) Duality Approach: We first define the Lagrangian function $\mathcal{L}$ associated with (18) as,

$$\mathcal{L}(\mathbf{W},\beta,\zeta) = MT_{\mathsf{TS}} \sum_{\forall(b,k)} W_k^b + \sum_{\forall(k,c)} \zeta^c \left(\mathbf{u}^c \mathbf{W}_k - W^{\mathsf{total}}\right) - \sum_{\forall b} \beta^b \left(\sum_{\forall k} z_k(W_k^b) - 1 + \bar{P}_{\mathsf{QoE}}\right), \tag{19}$$

where β^b and ζ^c are the Lagrangian multipliers; $\mathbf{u}^c$ stands for the vector generated from the c-th row of $\mathbf{U}$; and $\mathbf{W}_k = [W_k^1, W_k^2, ..., W_k^B]^T$. Then, the dual function of W_k^b can be defined as the minimum of the Lagrangian function as,

$$\mathsf{g}(\beta) = \min_{\mathbf{W}} \mathcal{L}(\mathbf{W},\boldsymbol{\beta},\boldsymbol{\zeta}) \text{ s.t. (17).} \tag{20}$$

To find the best lower bound that can be obtained from the Lagrange dual function, the dual problem can be written as,

$$\max_{\beta^b,\zeta^c} \mathsf{g}(\beta,\zeta) \text{ s.t. } \beta^b \geq 0, \zeta^c \geq 0. \tag{21}$$

Since problem (18) is convex, the dual-gap between the primary and dual problem is zero [44]. In the following, one will describe a searching approach to define the optimal solution. In particular, the dual problem is always convex, $\mathsf{g}(\beta^b,\zeta^c)$ can be maximized by using the standard sub-gradient method where the dual variables β^b and ζ^c are first initialized to random values in the dual feasibility region of $\beta^b > 0$, $\zeta^c > 0$, $\mathsf{g}(\beta^b,\zeta^c) > -\infty$ [45]. The dual variables can be iteratively updated as follows:

$$\beta^b_{[\ell+1]} = \left[\beta^b_{[\ell]} - \delta_{[\ell]}\left(\sum_{k=1}^{K} z_k(W_k^b) - 1 + \bar{P}_{\mathsf{QoE}}\right)\right]^+, \tag{22}$$

$$\text{and} \quad \zeta^c_{k,[\ell+1]} = \left[\zeta^c_{k,[\ell]} + \delta_{[\ell]}\left(\mathbf{u}^c \mathbf{W}_k - W^{\mathsf{total}}\right)\right]^+, \tag{23}$$

where the suffix $[\ell]$ represents the iteration index, $\delta_{[\ell]}$ is the step size, and $[x]^+$ is defined as $\max(0,x)$. This sub-gradient method guarantees the convergence for any initial primary point of $\{W_k^b\}$'s if the step-size $\delta_{[\ell]}$ is chosen appropriately so that $\delta_{[\ell]} \xrightarrow{\ell\to\infty} 0$ such as $\delta_{[\ell]} = 1/\sqrt{\ell}$ [44], [46].

2) Solving the optimization problem related to dual function: This section focuses on minimizing the Lagrangian function,

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W},\boldsymbol{\beta},\boldsymbol{\zeta}) \text{ s.t. constraint (17).} \tag{24}$$

Proposition 2. *The optimal solution of* (24) *is given as*

$$(W_k^b)^\star = \max(\alpha_k^b, \hat{W}_k^b), \tag{25}$$

where,

$$\hat{W}_k^b = \left\{\beta^b A_k^b (Q_{\mathsf{QoE}}+1) / \left[\left(MT_p + \sum_{c=1}^{C} \zeta^c U_{c,b}\right)T\right]\right\}^{1/(Q_{\mathsf{QoE}}+2)}, \tag{26}$$

A_k^b is as defined in (49) *and $U_{c,b}$ is the element on row c and column b of matrix* $\mathbf{U}$.

Proof: The proof is given in Appendix C. ∎

Algorithm 1 Duality-based Dynamic Capacity Allocation

```
Initialization:
    • Choose initial dual vector values $\beta^b$ and $\zeta^c$.
    • Select a tolerate $\epsilon$, step size $\delta$, and set $\Delta_1 = 1$, $\Delta_2 = 1$ $\ell = 0$.
    • Define $\mathbf{U}$ , $W^{\mathsf{total}}$ and provide values for $Q_{\mathsf{QoE}}$, $\bar{P}_{\mathsf{Blk}}$, $\bar{P}_{\mathsf{QoE}}$ and $Q_{\mathsf{max}}$.
while $\Delta_1 > \epsilon$ and $\Delta_2 > \epsilon$ do
    Given $\beta^b_{[\ell]}$ and $\zeta^c_{[\ell]}$, define $(W_k^b)^\star$'s as in (25) and (26).
    Based on $(W_k^b)^\star$'s, update $\beta^b_{[\ell+1]}$ as in (22).
    Based on $(W_k^b)^\star$'s, update $\zeta^c_{[\ell+1]}$ as in (23).
    Re-set $\Delta_1 := |\beta^b_{[\ell+1]} - \beta^b_{[\ell]}|$.
    Re-set $\Delta_2 := |\zeta^c_{[\ell+1]} - \zeta^c_{[\ell]}|$.
    Set $\ell := \ell + 1$.
end while
Return $\mathbf{W}_k^\star$.
```

Algorithm 2 Greedy-based Dynamic Capacity Allocation

```
Inputs:
    • Provide initial values for $\lambda^b_{i,k}$, $L_s$.
    • Provide values for $Q_{\mathsf{QoE}}$, $\bar{P}_{\mathsf{Blk}}$, $\bar{P}_{\mathsf{QoE}}$ and $Q_{\mathsf{max}}$.
for $b = 1$ to Number of beams do
    for $k = 1$ to Number of cycles do
        Calculate $\alpha_{k,1}, \alpha_{k,2}, \alpha_{k,3}$ .
        Calculate $\max(\alpha_{k,1}, \alpha_{k,2}, \alpha_{k,3})$.
    end for
end for
Return $\mathbf{W}^\star$.
```

3) Proposed Duality-based Algorithm: Thanks to the duality approach, the optimal solution of problem (18) can be obtained by alternatively solving problem (24) - the right-hand-side of (20) - as presented in Proposition 2, and updating Lagrangian multipliers β^b and ζ^c as in (22) and (23) in each iteration until the convergence. The optimization-based approach is summarized in Algorithm 1 where the iteratively solving process can be stopped when the gaps $\Delta_1 = \sum_{b=1}^{B} |\beta^b_{[\ell+1]} - \beta^b_{[\ell]}|$ and $\Delta_2 = \sum_{c=1}^{C} |\zeta^c_{[\ell+1]} - \zeta^c_{[\ell]}|$ are sufficiently small.

C. Greedy Algorithm

To mitigate the complexity of solving our problem given in the previous section, this section introduces a straightforward and efficient greedy algorithm. As evident, the primary challenge in addressing problem 16 arises from the coupling of all W_k^b throughout the entire time window T as illustrated in (16c) for each beam. In particular, directly handling this constraint needs to consider the average probability of QoE violation across the entire time window T. To simplify this process, we disregard the average value, prompting the system to meet the QoE requirement at every given moment. As such, constraint (16c) is replaced by a more stringent version, which is presented as follows:

$$\sum_{n=0}^{Q_{\mathsf{QoE}}} g_{n,b}(W_k^b, t) \geq 1 - \bar{P}_{\mathsf{QoE}}, \forall k, \forall b \text{ and } t \in \Omega_k. \tag{27}$$

This new constraint is similar to (16b). By employing the same approach handling (16b) given in Proposition 1, we first define the maximum arrival rate over one cycle as,

$$\lambda_{i,k}^b = \max_{t \in \Omega_k} \lambda_i^b(t). \tag{28}$$

Then, the new constraint in (27) can also be translated into,

$$W_k^b \geq \alpha_{k,3}^b = \max_{t \in \Omega_k} L g_{Q_{\mathsf{QoE}}}^{-1}\left(1 - \bar{P}_{\mathsf{QoE}}, t\right)/T_{\mathsf{TS}}. \tag{29}$$

Thanks to Proposition 1, we are able to estimate the required capacity of each beam over every cycle by taking the maximum capacity that satisfies the problem constraints. That is,

$$(W_k^b)^\star = \max\left(\alpha_{k,1}^b, \alpha_{k,2}^b, \alpha_{k,3}^b\right). \tag{30}$$

The greedy algorithm is summarized in Algorithm 2.

D. Other Benchmark Algorithm

This subsection introduces another benchmark solution for comparison purposes, which is developed by adapting the ISA given in [34]. The ISA is well-established for time-varying human resource management to satisfy the stochastic waiting time requirement of customers. In [34], the $M_t/M/S(t)$ queue model is employed where the serving capacity of one server (or employees) is fixed (M) while the number of servers (employees) can be varied over the time. Herein, $S(t)$ represents the number of allocated employees at time t, and the ISA is designed to determine $S(t)$ coping with the time-varying customer arrival rate efficiently.

As can be seen, the queuing model utilized in [34] is different from our $M_t/M_t/1$ scheme which is related to one server with variable serving capacity. Hence, in order to modify this work to address our problem, we assume that the capacity amount allocated to beam b in cycle k can be represented by a number of *"fixed capacity packages"*. Let W_0 (bps) be the capacity of one such package, and S_k^b denote the number of capacity packages assigned for beam b in cycle k. Then, we have,

$$W_k^b = S_k^b W_0. \tag{31}$$

In addition, the processing rate corresponding to one package can be estimated as

$$\mu_0 = W_0 T_{\mathsf{TS}}/L. \tag{32}$$

Algorithm 3 ISA-BASED CAPACITY ALLOCATION

```
Input:
    • Initialize a vector of the number of packages $S_k^b(0)$.
    • Set values of $\bar{P}_{\mathsf{Blk}}$ , $\bar{P}_{\mathsf{QoE}}$, $\mu_0$, $\Delta$ and $\Delta_0$, $\epsilon_1 = 10^{-4}$ , counter $i = 0$.
while $\Delta > \epsilon_1$ and $\Delta_0 > \epsilon_1$ do
    For every $t \in \Omega_k$ , calculate $\rho^b(t) = \lambda^b(t)/(\mu_0 S_k^b(i))$ and $\hat{P}_{n,b}(t)$ according to (14).
    Calculate $V_{\mathsf{bl},k}^b(i)$ as in (35) and $V_{\mathsf{QoE},k}^b(i)$ as in (38).
    Calculate $\psi_{\mathsf{bl},k}^b(i)$ as in (34) and $\psi_{\mathsf{QoE},k}^b(i)$ as in (37).
    Calculate $S_{\mathsf{bl},k}^b(i+1)$ as in (33) and $S_{\mathsf{QoE},k}^b(i+1)$ according to (36).
    Set $\Delta = |V_{\mathsf{QoE},k}^b(i) - \bar{P}_{\mathsf{QoE}}|$.
    Set $\Delta_0 = |V_{\mathsf{bl},k}^b(i) - \bar{P}_{\mathsf{Blk}}|$.
    Set $i = i + 1$.
    Set $S_k^b(i) = \max\left\{S_{\mathsf{bl},k}^b(i+1), S_{\mathsf{QoE},k}^b(i+1), S_k^b(0)\right\}$.
end while
Return $\mathbf{W}_k^b = L\mu_0 \mathbf{S}_k^b / T_{\mathsf{TS}}$.
```

Now, we can employ the ISA to optimize $\{S_k^b\}$'s by regarding the constraints (8b) and (8c) instead of the stochastic waiting time as designed in [34]. Specifically, at the initialization, $\{S_k^b\}$'s are randomly selected and then adjusted over iterations. There are two components to define S_k^b in every iteration i, the first is $S_{\mathsf{bl},k}^b(i)$ - being updated according to the blocking probability requirement, and the latter is $S_{\mathsf{QoE},k}^b(i)$ - being adjusted due to QoE-related demand. Here, constraint (8c) is considered for every cycle to update $S_{\mathsf{QoE},k}^b(i)$. To ensure compliance with the blocking probability requirement stated in (16b), $S_{\mathsf{bl},k}^b(i)$ is updated as follows:

$$S_{\mathsf{bl},k}^b(i+1) = \begin{cases} \lceil S_{\mathsf{bl},k}^b(i)\psi_{\mathsf{bl},k}^b(i) \rceil & \text{if } \psi_{\mathsf{bl},k}^b(i) \geq 1, \\ \lfloor S_{\mathsf{bl},k}^b(i)\psi_{\mathsf{bl},k}^b(i) \rfloor & \text{otherwise}, \forall k, \end{cases} \tag{33}$$

where $\lceil . \rceil$ and $\lfloor . \rfloor$ indicate the ceil and floor operators and $\psi_{\mathsf{bl},k}^b(i)$ is the blocking-probability influence factor corresponding to $S_k^b(i)$ and $\bar{P}_{\mathsf{Blk}}$. In particular, $\psi_{\mathsf{bl},k}^b(i)$ can be given as

$$\psi_{\mathsf{bl},k}^b(i) = 1 + \frac{V_{\mathsf{bl},k}^b(i) - \bar{P}_{\mathsf{Blk}}}{\bar{P}_{\mathsf{Blk}} i}, \forall k, \tag{34}$$

where $V_{\mathsf{bl},k}^b(i)$ indicates the maximum blocking probability during cycle k corresponding to $S_k^b(i)$. Specifically, $V_{\mathsf{bl},k}^b$ is given as follows:

$$V_{\mathsf{bl},k}^b = \max_{t \in \Omega_k} \sum_{n=0}^{Q_{\max}} \hat{P}_{n,b}(t)\Big|_{W_k^b = W_0 S_k^b(i)}. \tag{35}$$

Similarly, the $S_{\mathsf{QoE},k}^b(i)$ in iteration i is updated as follows:

$$S_{\mathsf{QoE},k}^b(i+1) = \begin{cases} \lceil S_{\mathsf{QoE},k}^b(i)\psi_{\mathsf{QoE},k}^b(i) \rceil & \text{if } \psi_{\mathsf{QoE},k}^b(i) \geq 1, \\ \lfloor S_{\mathsf{QoE},k}^b(i)\psi_{\mathsf{QoE},k}^b(i) \rfloor & \text{otherwise}, \forall k, \end{cases} \tag{36}$$

where $\psi_{\mathsf{QoE},k}^b(i)$ is a QoE-related influence factor corresponding to $S_k^b(i)$ and $\bar{P}_{\mathsf{QoE}}$. Here, $\psi_{\mathsf{QoE},k}^b(i)$ is expressed as,

$$\psi_{\mathsf{QoE},k}^b(i) = 1 + \frac{V_{\mathsf{QoE},k}^b(i) - \bar{P}_{\mathsf{QoE}}}{\bar{P}_{\mathsf{QoE}} i}, \forall k, \tag{37}$$

where $V_{\mathsf{QoE},k}^b(i)$ indicates the maximum probability of violating the target QoE requirement during cycle k with $S_k^b(i)$ as

$$V_{\mathsf{QoE},k}^b = \max_{t \in \Omega_k} \sum_{n=0}^{Q_{\mathsf{QoE}}} \hat{P}_{n,b}(t)\Big|_{W_k^b = W_0 S_k^b(i)}. \tag{38}$$

In addition, to meet the condition $\rho^b(t) \le 1$ of (16c), it is essential to carefully set S_k^b in a way $\mu_0 S_k^b \ge \lambda^b(t)$ for every $t \in \Omega_k$. Then, S_k^b can be updated as

$$S_k^b(i) = \max\left\{S_{\mathsf{bl},k}^b(i+1), S_{\mathsf{QoE},k}^b(i+1), S_k^b(0)\right\} \tag{39}$$

Accordingly, the adapted ISA is summarized in Algorithm 3.

E. Complexity Analysis

The complexity of Algorithm 1 arises from calculating α_k^b and also processing a number of loops in each of that $\beta_{[\ell]}^b$, $\zeta_{[\ell]}^c$, and $(W_k^b)^\star$ are estimated as given in (22), (23), (26), respectively. As given in (17), the complexity of estimating α_k^b is the order of $O\left(KBMQ_{\mathsf{max}}^2\right)$. Regarding the effort of estimating $z_k(W_k^b)$ and $\sum_{k=1}^K z_k(W_k^b)$, the complexity due to equation (22) can be given as $O\left(K^2BMQ_{\mathsf{QoE}}\right)$. Similarly, the complexity due to equation (23) is the order of $O(KBC)$. Next, the complexity due to equation (26) is associated with the summation of C elements and a power-of-Q_{QoE} calculator. Hence, the computation effort for calculating $\left\{\hat{W}_k^b\right\}$'s corresponding to B beams and K cycles can be the order of $O\left(KB(C+Q_{\mathsf{QoE}})\right)$. One assumes that implementing Algorithm 1 required $\ell^{(1)}$ iterations to get convergence and obtain the solution, the overall complexity of the algorithm taking the highest degree polynomial becomes

$$X^{\mathsf{Alg.1}} = \\ O\left(KB\left[MQ_{\mathsf{max}}^2 + \ell^{(1)}(KMQ_{\mathsf{QoE}} + 2C + Q_{\mathsf{QoE}})\right]\right). \tag{40}$$

This shows that our problem can be easily solved and converged in a polynomial amount of computation time. Considering the greedy algorithm, one can see that the outcome can be obtained by estimating $\alpha_{k,1}, \alpha_{k,2}, \alpha_{k,3}$. According to (17) and (29), the required computation effort for implementing Algorithm 2 can be described as

$$X^{\mathsf{Alg.2}} = O\left(KBM(Q_{\mathsf{max}}^2 + Q_{\mathsf{QoE}}^2)\right). \tag{41}$$

Consequently, we study the complexity of the ISA method. As summarized in Algorithm 3, the process of ISA method encompasses multiple loops of estimating S_k^b through equations (33) and (36). In each loop, the heaviest task of determining S_k^b replies on calculating $V_{\mathsf{bl},k}^b$ and $V_{\mathsf{QoE},k}^b$ in (35) and (38), respectively. Therefore, the complexity of ISA can be given as

$$X^{\mathsf{Alg.3}} = O\left(\ell^{(3)}KBM(Q_{\mathsf{max}}^2 + Q_{\mathsf{QoE}}^2)\right). \tag{42}$$

This has shown a relatively low computation effort. While both the greedy and ISA approaches are simpler than the duality method, we favored the latter because of its superior efficiency, as showcased in Section V-B. Moreover, the duality method, unlike some algorithms which may require exponential time to converge, promises convergence in polynomial time. This not only ensures more predictable computational demands but also bolsters its viability as an optimal approach for real-world applications.

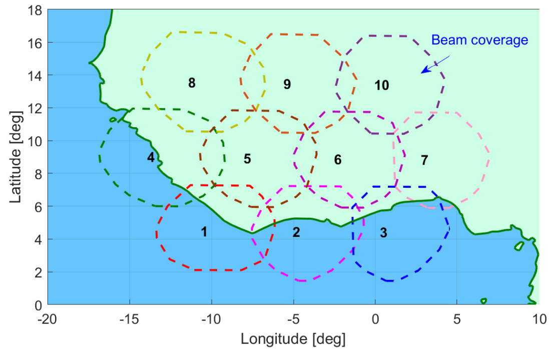


Fig. 5: Considered GEO beam footprint pattern with $N = 10$.

TABLE II: Simulation parameters.

Parameters	**Considered values**
Cycle duration (M)	10 minutes
Normalized packet length (L)	64 (KBytes) [47]
Maximum buffer size (Q_{max})	30 packets 2 MB [48]
Total available capacity of the satellite (Gbps)	2.1 Gbps [49]
Number of beams (J) per cluster	3 [50]
Number of virtual beams (N)	10 [51]
Number of cycles (K)	6
Number of time slots	180000
Price per Mbits (γ)	0.1 Euros [52]
Random (r_s^b)	[1 – 4]
Random (a_s^b)	[0 – 1]
Random phase (ϕ_s^b)	[1 – 360] degrees
Target blocking probability ($\bar{P}_{\mathsf{Blk}}$)	0.01 [53]
Time slot duration (T_{TS})	20 ms [54]

V. Performance Evaluation and Numerical Results

In this section, we simulate and analyze a time-varying queuing model to estimate the stochastic blocking probability over time and to find the optimal capacity that can satisfy the defined QoE and blocking probability requirements.

A. Simulation Setup and Parameters

In this subsection, we conduct a Monte Carlo simulation consisting of 5000 independent data trials. The simulation includes generating random arrival rates based on a time-varying Poisson process [39]–[41] for various time slots, as well as assigning time-varying service rate values for different cycles. In each iteration, the arrival rate function is chosen to represent time-varying demand that varies between zero and the assumed system's maximum capacity, using a sinusoidal representation as described in [55]. Three data flows are generated for every beam and the corresponding time-varying arrival-rate functions for beam b are given as

$$\lambda_i^b(t) = r_i^b(1 + \sin(a_i^b t + \phi_i^b)), \tag{43}$$

where suffix i stands for data flow i and r_i^b is an influencing factor corresponding to the average number of arriving packets at time t, a_i^b is a positive number influencing the periodicity of the arrival time of packets, and ϕ_i^b's are phase shift angles. Here, r_i^b is selected randomly in a range of $[1, 4]$ so that the average of the total demand $(\sum_{s=1}^S L\lambda_s^b(t)/T_{\mathsf{TS}})$ is not greater than the assumed beam capacity of 700 Mbps as presented in [49]; a_i^b is selected randomly a range of $(0, 1]$ as in [56],

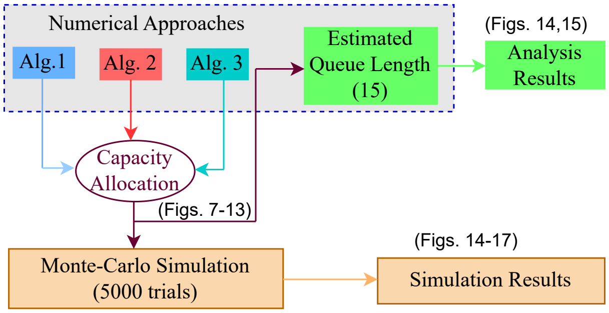


Fig. 6: The diagram of obtaining numerical, analysis, and simulation results.

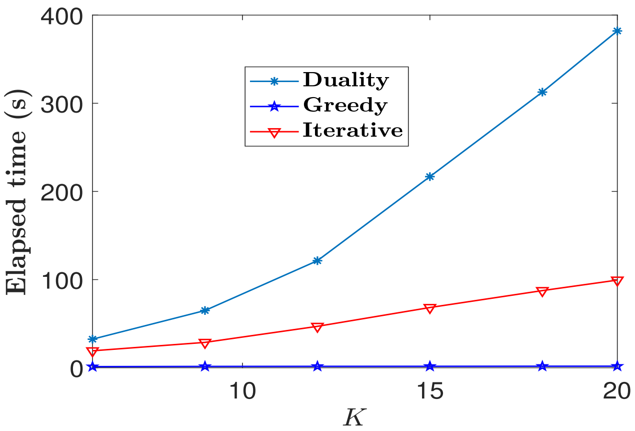


Fig. 7: Running time versus the number of cycles (K).

[57], while ϕ_i^b is chosen randomly within $[0, 360]$ to ensure all flows have different peak and off-peak periods. The remaining parameters provided in Table II are adopted for all simulations unless specified otherwise.

By utilizing the mean arrival rate function obtained from the Monte Carlo trials and the service rate values obtained from the optimal allocated capacity according to (5), we compute the queue length, blocking probability, and the probability of violating QoE requirements for each time slot t. The simulation results as compared to the analytical results are explained in the next sub-section for different values of the considered parameters. Regarding the renting cost function $f(W)$, we exploit a linear form as $f(W) = \gamma W$ [7], [58] where γ represents the price per capacity unit, i.e., Euros/Mbits. Additionally, Fig. 6 illustrates the diagram of obtaining numerical, analysis, and simulation results in this section. As can be seen, Algorithms 1-3 are first employed to determine the rented capacity solutions based on which the costs can be calculated. They are so-called numerical results which are demonstrated in Figs. 7-13. Furthermore, the capacity outcomes are utilized to obtain the analysis results by using (15), which are illustrated in Figs. 14 and 15. Additionally, the Monte Carlo simulation results based on the numerical capacity solutions are illustrated in Figs. 14-17.

B. Numerical Results and Discussion

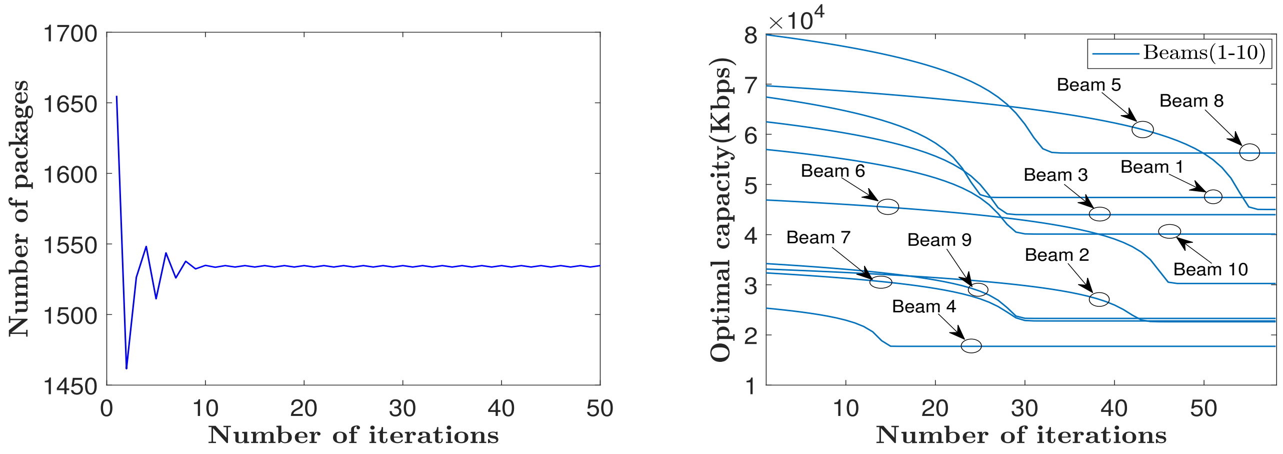
This section first shows the running time and convergence of the proposed algorithm, then we investigate the effect of varying parameters, namely Q_{QoE}, $\bar{P}_{\mathsf{Blk}}$, Q_{max}, $\bar{P}_{\mathsf{QoE}}$, and K, on the optimal allocated capacity and total renting cost to meet the time-varying demand. Assuming both polarizations are used in all beams, we can put a minimum of 3 adjacent beams per cluster to avoid cross-beam interference [50]. Hence, our numerical results are based on 3-beam clustering framework arrangement with 10-beam footprints as described in Fig. 5. In particular, there are 10 clusters in this beam pattern setting which are $(1,4,5)$; $(1,2,5)$; $(2,5,6)$; $(2,3,6)$; $(3,6,7)$; $(4,5,8)$; $(5,8,9)$; $(5,6,9)$; $(6,9,10)$ and $(6,7,10)$. Accordingly, constraint (2) corresponding to cycle k is demonstrated as

$$\begin{pmatrix} 1&0&0&1&1&0&0&0&0&0\\ 1&1&0&0&1&0&0&0&0&0\\ 0&1&0&0&1&1&0&0&0&0\\ 0&1&1&0&0&1&0&0&0&0\\ 0&0&1&0&0&1&1&0&0&0\\ 0&0&0&1&1&0&0&1&0&0\\ 0&0&0&0&1&0&0&1&1&0\\ 0&0&0&0&1&1&0&0&1&0\\ 0&0&0&0&0&1&0&0&1&1\\ 0&0&0&0&0&1&1&0&0&1 \end{pmatrix} \begin{pmatrix} W_k^1\\ W_k^2\\ W_k^3\\ W_k^4\\ W_k^5\\ W_k^6\\ W_k^7\\ W_k^8\\ W_k^9\\ W_k^{10} \end{pmatrix} \leq W^{\mathsf{total}} \begin{pmatrix} 1\\1\\1\\1\\1\\1\\1\\1\\1\\1 \end{pmatrix}. \tag{44}$$

Fig. 7 shows the running time of duality, greedy and ISA algorithms with varying values of K, using Matlab (tic-toc method) on an Intel(R) Core(TM) i7 processor. The plot for the duality algorithm indicates that it is close to the expected polynomial time complexity which is proportional to power 2 of the size of cycles, which confirms the analysis results given in Section IV-E. Next, we present the convergence plot of our proposed algorithm alongside the benchmark ISA in Fig.8. The figure in 8b illustrates the changes in capacities assigned to the beams when Algorithm 1 is implemented. As shown, the assigned capacity for each beam decreases before stabilizing at its minimum value across iterations. Moreover, different beams necessitate various numbers of iterations to reach convergent capacity values. Similarly, the fluctuations in the total number of packages allocated to all beams in Fig. 8a are the result of employing Algorithm 3 across iterations. The consistent increment or decrement by 1 package in the plot arises from the rounding effects in the expression (33) of Algorithm 3.

Figs. 9a and 10a show the total allocated capacity per cycle $(\sum_{b=0}^{B} W_k^b)$ and the total capacity rental cost $(\sum_{b=0}^{B}\sum_{k=0}^{K}\gamma M T_{\mathsf{TS}} W_k^b)$ obtained by using the Lagrangian duality, ISA and greedy algorithms at varying Q_{QoE} values, respectively. The parameters $Q_{\mathsf{max}} = 32$, $\bar{P}_{\mathsf{Blk}} = 0.01$, $\bar{P}_{\mathsf{QoE}} = 0.05$, $K = 12$, and $T_{\mathsf{TS}} = 20$ ms have been taken into account. Here, one assumes that the capacity of one package is set to 5 Mbps for implementing ISA algorithm, which is equivalent to the scenario of the transmission over a 1 MHz sub-channel with 32-QAM modulation at acceptable signal-to-noise ratio [59]. The study results indicate that a system catering to users with a higher tolerance for waiting requires less capacity and thus incurs lower costs. Moreover, higher values of Q_{QoE} and Q_{max} can lead to lower required capacity, resulting in smaller costs.

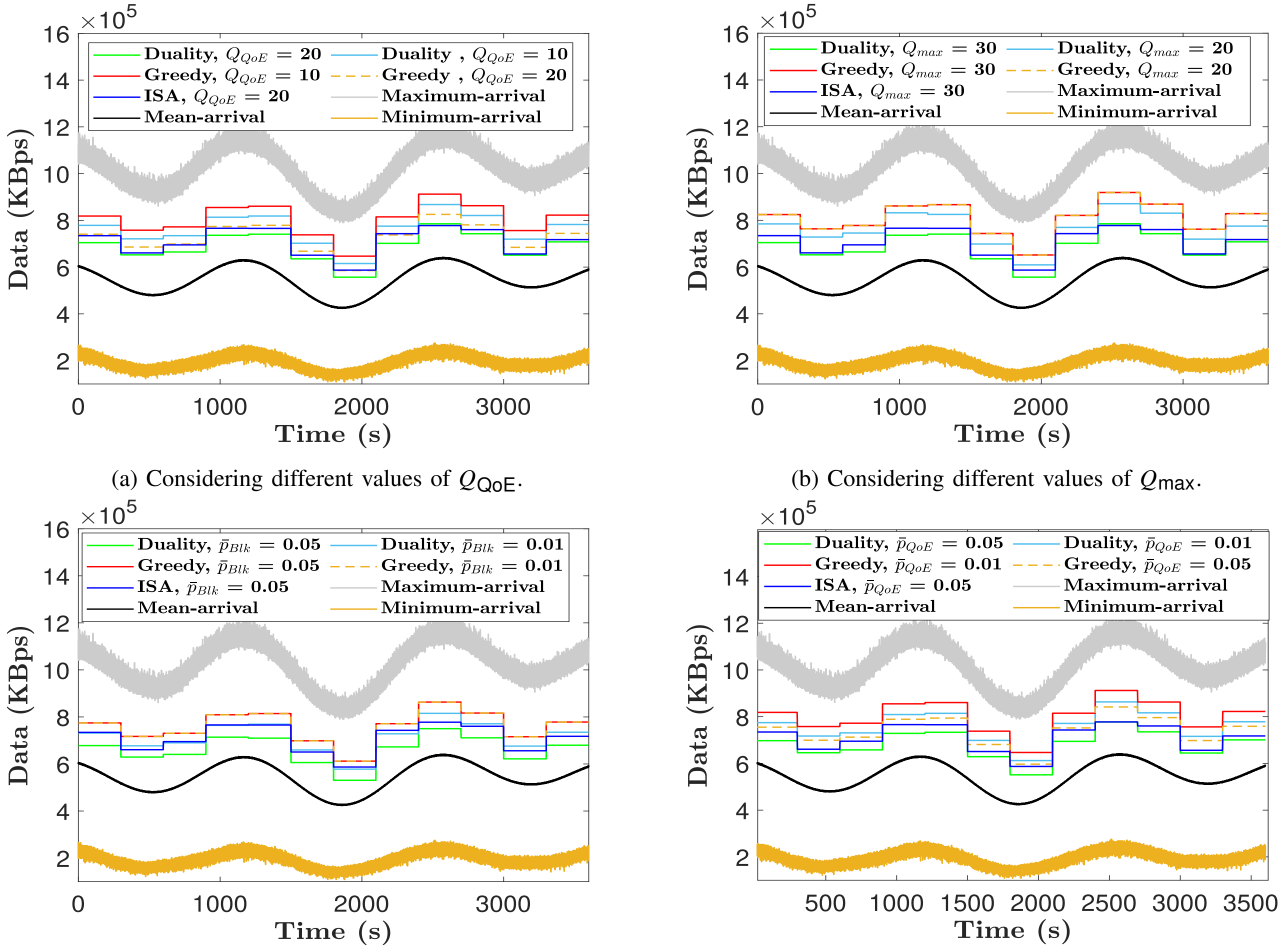
Figs. 9b and 10b show the sum of optimal allocated capacity per cycle and total rental cost of all beams, respectively, when



(a) ISA.

(b) Lagrangian duality.

Fig. 8: Convergence plot.



(a) Considering different values of Q_{QoE}.

(b) Considering different values of $Q_{\max}$.

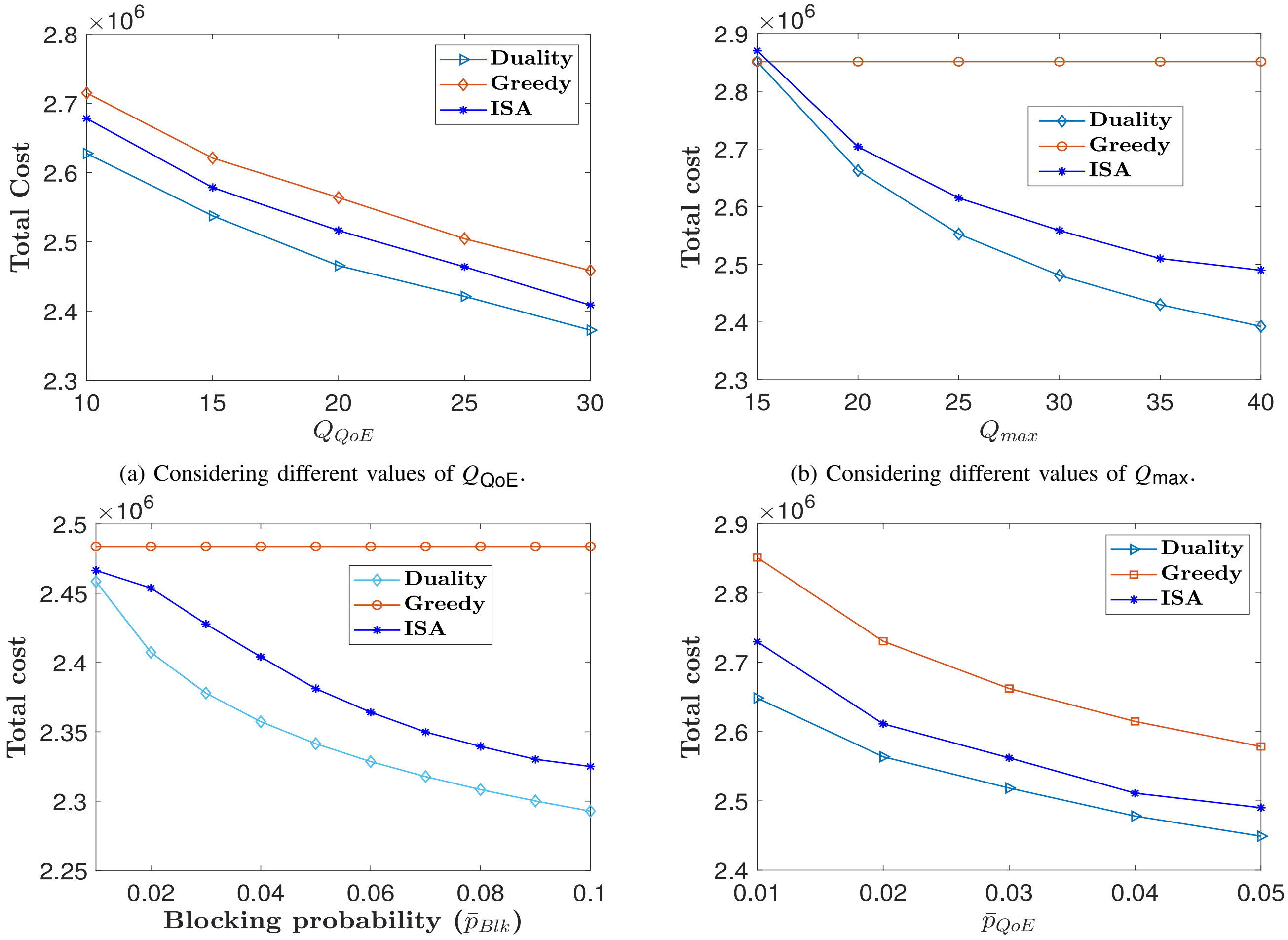
(c) Considering different values of $\bar{P}_{\mathsf{Blk}}$.

(d) Considering different values of $\bar{P}_{\mathsf{QoE}}$

Fig. 9: Optimal allocated capacity to satisfy the time-varying arrivals (demand).

varying $Q_{\max}$ while keeping other parameters ($Q_{\mathsf{QoE}} = 15$, $K = 12$, $\bar{P}_{\mathsf{Blk}} = 0.01$, and $\bar{P}_{\mathsf{QoE}} = 0.05$) constant. The results obtained through the Lagrangian duality and ISA method indicate that allowing more packets stored during congestion requires less allocated capacity. In contrast, the greedy algorithm assumes the maximum capacity required to meet the target Q_{QoE} requirement, neglecting the buffer size and its influence on capacity allocation. The figures also reveal that a larger buffer size and greater queuing delay tolerance result in lower allocated capacity and rental costs.

Fig. 9c demonstrates the allocated capacity per cycle at various $\bar{P}_{\mathsf{Blk}}$ values using the Lagrangian duality, ISA, and



(a) Considering different values of Q_{QoE}.

(b) Considering different values of $Q_{\max}$.

(c) Considering different values of $\bar{P}_{\text{Blk}}$.

(d) Considering different values of $\bar{P}_{\text{QoE}}$.

Fig. 10: Total renting cost of the optimal allocated capacity to all beams.

greedy algorithms at $Q_{\max} = 32$, $Q_{\text{QoE}} = 20$, $\bar{P}_{\text{QoE}} = 0.05$, and $K = 12$. The Lagrangian duality and ISA approaches show that systems with lower blocking probability requirements have greater allocated capacities than those with higher blocking probabilities. However, the greedy algorithm provides the same optimal capacity for all $\bar{P}_{\text{Blk}}$ values, as it only considers the maximum value associated with queuing delay violations and not with blocking probability. This makes our model more efficient in accounting for blocking probability. Fig. 10c shows the relationship between blocking probability and total rental costs at different Q_{QoE} values. As can be seen, the lower blocking probability requirements return higher allocated capacities and costs. For example, based on the obtained results, an increase in $\bar{P}_{\text{Blk}}$ from 0.01 to 0.05 results in a cost reduction of approximately 5%, while an increase from 0.01 to 0.1 leads to a reduction of 7.23%.

Next, Figs. 9d and 10d depict the relationship between $\bar{P}_{\text{QoE}}$ and total allocated capacity as well as rental costs obtained by implementing the three algorithms. As expected, the outcomes of all three algorithms imply that a smaller probability of violating the QoE requirement necessitates the SP to allocate a higher capacity and, conversely, less capacity for a higher probability of violation. For instance, the obtained result indicates, an increase in $\bar{P}_{\text{QoE}}$ from 0.01 to 0.05 results in a reduction of the renting cost by 6.11%.

Figs. 11 and 12 present the total allocated capacities per cycle and the associated renting costs for the three algorithms for various values of K. From the figures, it's evident that when K rises, there's a decrease in the optimal capacity allocation. This trend suggests that a more adaptive system can fulfill demand using less capacity, leading to reduced costs. For instance, results show that when K increases from 6 to 18, the renting cost drops by 11.38%, and a surge from 12 to 18 results in a decline of 6.1%. Parameters for this analysis, including $Q_{\max} = 32$, $Q_{\text{QoE}} = 20$, $\bar{P}_{\text{Blk}} = 0.01$, and $\bar{P}_{\text{QoE}} = 0.05$, were consistently considered.

The results, as depicted in all the above figures, highlight the superiority of the duality method over both ISA and greedy algorithms in terms of flexibility and adaptiveness. For instance based on the obtained result and for the case where $Q_{\max} = 32$, $Q_{\text{QoE}} = 20$, $\bar{P}_{\text{Blk}} = 0.01$, $\bar{P}_{\text{QoE}} = 0.05$ and $K = 6$, the proposed model can meet the requirement at a 9.85% and 3.1% lower cost compared to the greedy and ISA algorithms. However, as K increases, the greedy algorithm becomes as efficient as the proposed method, as larger values of K represent nearly immediate capacity changes, which are ideal conditions for

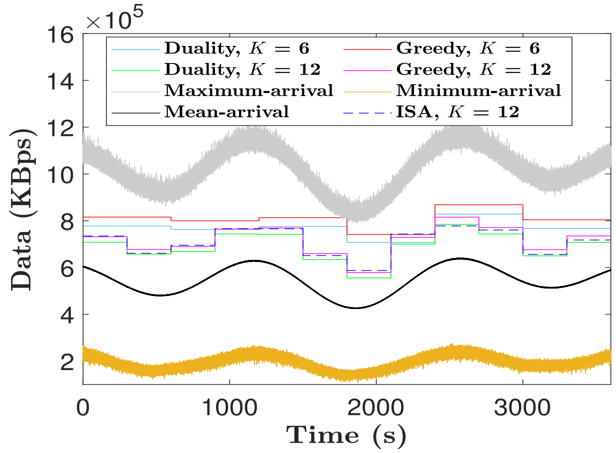


Fig. 11: Sum of the optimal allocated capacity of all beams at different values of K.

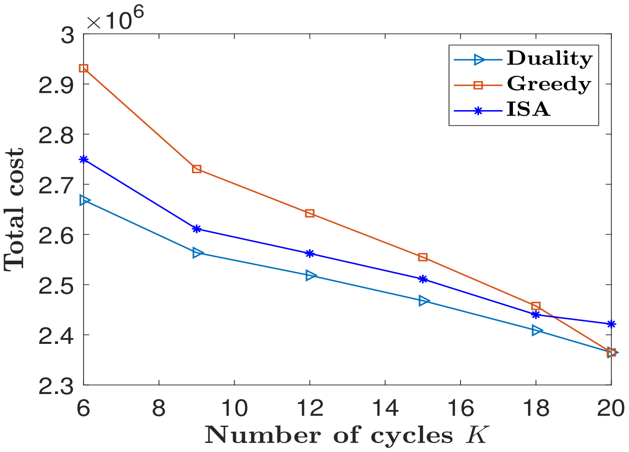


Fig. 12: Total renting cost versus K.

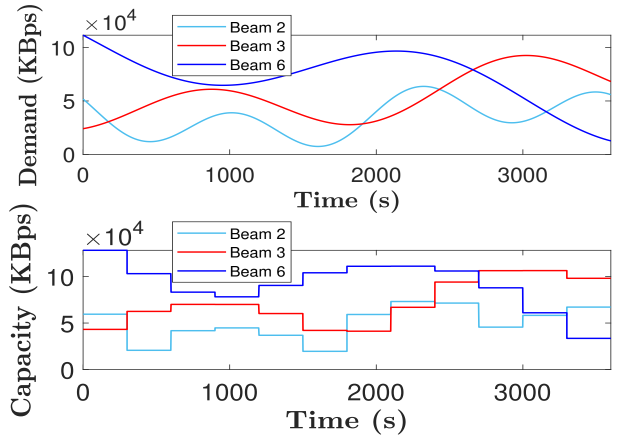


Fig. 13: Optimal capacity allocation to beams in the same cluster for different arrival rates.

greedy algorithms to perform well.

We need to examine if the capacity is shared among beams based on corresponding demand. We also need to assess whether the proposed model meets the requirements discussed in Section II-C. Fig. 13 displays the demand per beam as a function of the mean arrival rate ($L\lambda^b(t)/T_{\mathsf{TS}}$) and the optimal allocated capacity to different beams in a random cluster consisting of beams 2, 3, and 6. In every cycle, a higher capacity is assigned to the beam with the highest arrival rate, which corresponds to the highest demand, as demonstrated in the figure. This allocation meets the requirements in equation (2). For this demonstration, the parameters used are $Q_{\mathsf{max}} = 32$, $Q_{\mathsf{QoE}} = 20$, $\bar{P}_{\mathsf{Blk}} = 0.01$, $\bar{P}_{\mathsf{QoE}} = 0.05$, and $K = 12$.

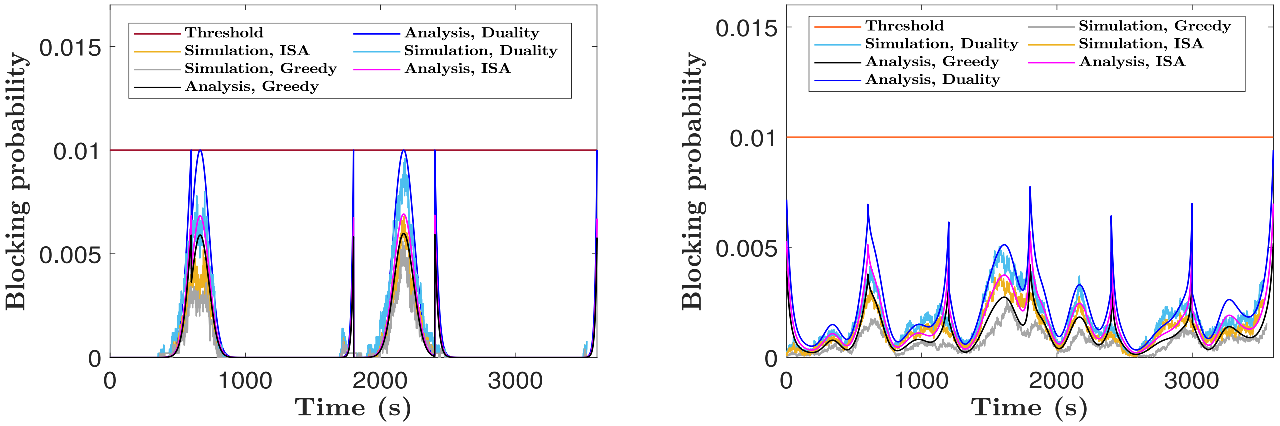
TABLE III: $\bar{P}_{\mathsf{QoE}}$ values at $Q \geq Q_{\mathsf{QoE}} = 20$ threshold using optimal capacity.

	Q_{QoE}	**20**	**24**	**28**	**32**
Duality	Analysis	0.01	0.0066	0.0025	0.0016
	Simulation	0.0093	0.0061	0.0021	0.0014
Greedy	Analysis	0.0089	0.0052	0.0019	0.0011
	Simulation	0.0083	0.0048	0.0017	0.0010
ISA	Analysis	0.009	0.0062	0.0020	0.0013
	Simulation	0.0086	0.0059	0.0018	0.0012

The Figs. 14a, 14b and 15 depict the mean blocking probability over time for a randomly selected beam, the average of the mean blocking probability of all beams, and the mean QoE requirement violation probability respectively. The target blocking probability of $\bar{P}_{\mathsf{Blk}} = 0.01$ was set with the parameters $Q_{\mathsf{max}} = 32$, $Q_{\mathsf{QoE}} = 20$, $\bar{P}_{\mathsf{QoE}} = 0.05$, and $K = 6$. The results show that all techniques satisfy the blocking probability requirement. Figures 14a, 14b, 15, and Table III clearly illustrate the close alignment between our analytical and simulation methods. Furthermore, the minor differences in average blocking probability and queue length across various beams, as depicted in Figs. 16 and 17, affirm the accuracy and validity of our proposed method. This shows an effective integration of the inherent randomness and stochastic nature of user traffic demands for capacity management. Additionally, Table III shows that the queuing delay requirement is duly met across all instances of Q surpassing the threshold Q_{QoE}. Interestingly, the greedy and ISA algorithms result in lower blocking and QoE violation probabilities than the proposed duality method. Similarly, the greedy and ISA algorithms result in lower queuing length than the proposed duality method as shown in Fig. 17. This is because of that these benchmarks return the higher assigned capacity for beams.

C. Discussion on Feasibility of Practical Implementation

The simulation results underscore the computational efficiency of the proposed method, demonstrating its capability to adeptly manage the dynamic and fluctuating traffic demands inherent to SatCom systems. Specifically, as illustrated in Fig. 7, the method takes mere minutes (under 100 seconds when $K \leq 12$) to determine the optimal amount of rented BW over a one-hour time window. It's pertinent to highlight that this execution time can be further trimmed when the proposed algorithm runs on a more powerful industrial-grade computer. Such a run-time is practically viable, allowing the SPs to ascertain the necessary capacity prior to entering rental agreements with the IPs. Another pivotal factor for the successful implementation of our proposed algorithm in the practical systems is a deep understanding and accurate estimation of the time-varying arrival rate function, specifically, $\left\{\lambda_s^b(t)\right\}$'s. However, within the scope of this study, one does not delve into traffic model estimation. Several existing studies,



(a) Blocking probability over time for a random beam.

(b) Average blocking probability for all beams.

Fig. 14: Blocking probability achieved by analysis and simulation.

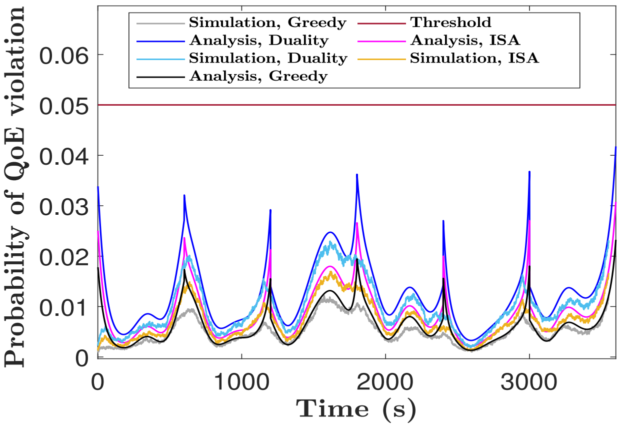


Fig. 15: Average probability of QoE violation for all beams.

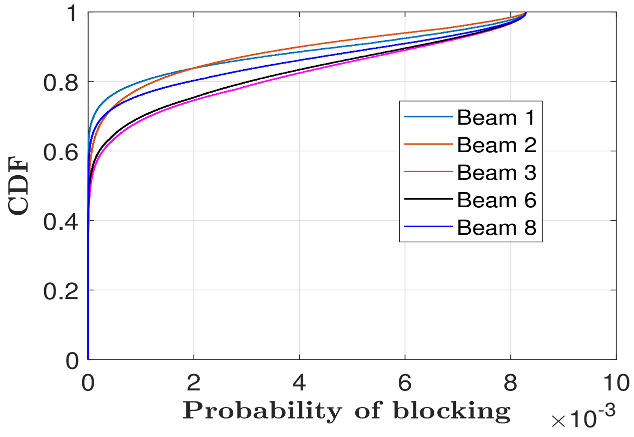


Fig. 16: CDF of blocking probabilities over time-window of beams $1, 2, 3, 6, 8$.

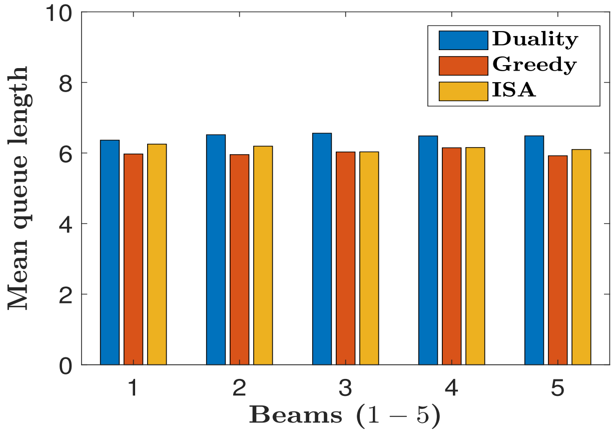


Fig. 17: Mean value of queue length over time-window T and 5000 trials due to beams $1-5$.

including those by [60]–[62], have dedicated efforts to unpack this intricate domain. Their insights suggest that a machine-learning-based model, which adjusts based on real-time data and historical patterns, could be the most efficient way to determine the stochastic information of the network traffic. The prospect of capacity management informed by traffic flow predictions is intriguing and is a topic we intend to address in future studies.

VI. Conclusion

In conclusion, this paper has proposed a novel and efficient dynamic capacity allocation model for multi-beam GEO satellite systems. The method aims to minimize the renting cost while ensuring the target blocking probability and QoE violation requirements. Traffic arrivals are modeled using the $M_t/M_t/1$ queueing model, and the stochastic queue length was estimated using the CTMC. The optimization problem has been solved using the Lagrangian duality method and the obtained results demonstrate its effectiveness and superiority over the benchmark ISA and greedy algorithms. Future work can include the extension of this model to a more complex network architecture, for example, the integration of the proposed model with terrestrial networks for a more comprehensive solution. Additionally, exploring the potential of machine learning techniques to further optimize the capacity allocation process can also be of great interest. Furthermore, considering the impact of flow prioritization and different traffic characteristics on the proposed model can provide deeper insights into the system's behavior and enhance its practicality.

APPENDIX A
PROOF OF PROPOSITION 1

Proof: Denote $\mu^b(t) = \mu_k^b$ in cycle k-th. We also have $\rho^b(t) = \lambda^b(t)/\mu_k^b$, $\forall t \in \Omega_k$ and $\mu^b(t) = \frac{W^b(t)T_{\mathsf{TS}}}{L}$. Then, constraint (16d) can be transferred into the following requirement

$$W_k^b \geq \alpha_{k,1}^b = \max_{t \in \Omega_k} L\lambda^b(t)/T_{\mathsf{TS}}. \tag{45}$$

Similarly, the constraint (16b) will be equivalent to

$$W_k^b \geq \alpha_{k,2}^b = \max_{t \in \Omega_k} Lg_{Q_{\max}}^{-1}(1 - \bar{P}_{\mathsf{Blk}}, t)/T_{\mathsf{TS}}. \tag{46}$$

The results given in (45) and (46) yield the lower bound of W_k^b as $\alpha_k^b = \max\left(\alpha_{k,1}^b, \alpha_{k,2}^b\right)$, which has completed the proof of Proposition 1. ■

APPENDIX B
PROOF OF THEOREM 1

Proof: Let $y_{Q_{\mathsf{QoE}}}(W_k^b, t) = \sum_{n=0}^{Q_{\mathsf{QoE}}} g_{n,b}(W_k^b, t)$. Taking $Y^b(t) = L\lambda^b(t)/T_{\mathsf{TS}}$, one can express $y_{Q_{\mathsf{QoE}}}(W_k^b, t)$ as,

$$y_{Q_{\mathsf{QoE}}}(W_k^b, t) = 1 - \left(Y^b(t)/W_k^b\right)^{Q_{\mathsf{QoE}}+1}. \tag{47}$$

Hence, we can see that $y_{Q_{\mathsf{QoE}}}(W_k, t)$ is a concave function with respect to W_k for any value of $\lambda^b(t)$ that satisfies $\lambda^b(t)/\mu_k^b < 1$. Using notation $y_{Q_{\mathsf{QoE}}}(W_k, t)$, we further denote $z_k(x) = \frac{1}{T}\int_{kM}^{(k+1)M} y_{Q_{\mathsf{QoE}}}(x, t)dt$. Again taking the integral with respect to t, we have

$$z_k(W_k^b) = MT_{\mathsf{TS}}/T - A_k^b/\left(T(W_k^b)^{Q_{\mathsf{QoE}}+1}\right), \tag{48}$$

where

$$A_k^b = \int_{(k-1)M}^{kM} Y^b(t)^{Q_{\mathsf{QoE}}+1}dt. \tag{49}$$

Similar to $y_{Q_{\mathsf{QoE}}}(W_k^b, t)$, $z_k(W_k^b)$ is also a concave function of W_k. Then, constraint (16c) can be rewritten as $\sum_{\forall k} z_k(W_k^b) \geq 1 - \bar{P}_{\mathsf{QoE}}$, since it is in the form of a concave function greater than a constant, it must be convex. Constraints (16b) and (16d) that are merged to (17) and equation (2) are also linear constraints. Hence, problem (16) must be convex. ■

APPENDIX C
PROOF OF PROPOSITION 2

Proof: As can be observed, the minimum value of $\mathcal{L}(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\zeta})$ can be defined by equating the partial derivative of $\mathcal{L}(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\zeta})$ with respect to W_k^b to zero, i.e.,

$$\frac{\partial \mathcal{L}(\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\zeta})}{\partial W_k^b} = MT_p - \frac{\beta^b A_k^b (Q_{\mathsf{QoE}} + 1)}{T(W_k^b)^{Q_{\mathsf{QoE}}+2}} + \sum_{\forall c} \zeta^c U_{c,b} = 0. \tag{50}$$

The solution of this equation can be described as

$$\hat{W}_k^b = \left\{\beta^b A_k^b (Q_{\mathsf{QoE}} + 1)/\left[\left(MT_p + \sum_{c=1}^{C} \zeta^c U_{c,b}\right)T\right]\right\}^{1/(Q_{\mathsf{QoE}}+2)}. \tag{51}$$

Then, by considering constraint (17), the optimal value of W_k^b can be expressed as in (25), which completes the proof of Proposition 2. ■

REFERENCES

[1] T. Mezgebo, V. N. Ha, E. Lagunas, J. Grozt, and S. Chatzinotas, "Qoe-oriented resource allocation design coping with time-varying demands in wireless communication networks," in *IEEE 96th Vehicular Technology Conference, (VTC-Fall 2022)*, 2022, pp. 1–6.

[2] F. Rinaldi, H.-L. Maattanen, J. Torsner, S. Pizzi, S. D. Andreev, A. Iera, Y. Koucheryavy, and G. Araniti, "Non-terrestrial networks in 5g & beyond: A survey," *IEEE Access*, vol. 8, pp. 165 178–165 200, 2020.

[3] G. Fontanesi, F. Ortíz, E. Lagunas, V. M. Baeza, M. Vázquez, J. A. Vásquez-Peralvo, M. Minardi, V. N. Ha, P. J. Honnaiah, C. Lacoste, Y. Drif, T. S. Abdu, G. Eappen, J. Rehman, L. M. Garcés-Socorrás, W. A. Martins, P. Henarejos, H. Al-Hraishawi, J. C. M. Duncan, T. X. Vu, and S. Chatzinotas, "Artificial intelligence for satellite communication and non-terrestrial networks: A survey," 2023.

[4] V. N. Ha, E. Lagunas, T. S. Abdu, H. Chaker, S. Chatzinotas, and J. Grotz, "Large-scale beam placement and resource allocation design for MEO-constellation SATCOM," in *IEEE ICC Workshop - 6GSatComNet*, 2023.

[5] eutelsat, "Satellite-as-a-Service: The Future of Satellite Network Communications," https://www.eutelsat.com/fr/home/news--resources/blog/satellite-as-a-service---the-future-of-connectivit.html, 2022, [Online; accessed 14-OCT-2022].

[6] R. Schradin, "The Government Satellite Report:Why Satellite as a Service is the Future of Government SATCOM," https://www.eutelsat.com/en/blog/satellite-as-a-servicethefutureofsatellitenetworkcommunications.html, 2023, "[Online; accessed 8-March-2023].

[7] S. Soursos, C. A. Courcoubetis, and R. R. Weber, "Dynamic bandwidth pricing: Provision cost, market size, effective bandwidths and price games," *J. Univers. Comput. Sci.*, vol. 14, pp. 766–785, 2008.

[8] W. Zhang, Y. Xue, J. Wu, and X. Xu, "Satellite as a service: a hybrid resource management framework for space-terrestrial integrated networks," in *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, 2020, pp. 171–174.

[9] M. Höyhtyä, S. Boumard, A. Yastrebova, P. Järvensivu, M. Kiviranta, and A. Anttonen, "Sustainable satellite communications in the 6g era: A european view for multilayer systems and space safety," *IEEE Access*, vol. 10, pp. 99 973–100 005, 2022.

[10] H. Du, J. Liu, D. Niyato, J. Kang, Z. Xiong, J. Zhang, and D. I. Kim, "Attention-aware resource allocation and qoe analysis for metaverse xurllc services," *IEEE Journal on Selected Areas in Communications*, 2023.

[11] Y. Xu, F. Yin, W. Xu, J. Lin, and S. Cui, "Wireless traffic prediction with scalable gaussian process: Framework, algorithms, and verification," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1291–1306, 2019.

[12] P. Jubba Honnaiah, "Demand-based optimization for adaptive multi-beam satellite communication systems," Ph.D. dissertation, University of Luxembourg Luxembourg, Luxembourg, 2022.

[13] N. Mazzali, M. R. Bhavani Shankar, and B. Ottersten, "On-board signal predistortion for digital transparent satellites," in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2015, pp. 535–539.

[14] M. Vincenzi, E. Lopez-Aguilera, and E. Garcia-Villegas, "Maximizing infrastructure providers' revenue through network slicing in 5g," *IEEE access*, vol. 7, pp. 128 283–128 297, 2019.

[15] V. N. Ha and L. B. Le, "End-to-end network slicing in virtualized ofdma-based cloud radio access networks," *IEEE Access*, vol. 5, pp. 18 675–18 691, 2017.

[16] C. L. G. Batista, F. Mattiello-Francisco, and A. Pataricza, "Heterogeneous federated cubesat system: problems, constraints and capabilities," 2022. [Online]. Available: https://arxiv.org/abs/2203.14721

[17] I. Sousa, M. P. Queluz, and A. Rodrigues, "A survey on qoe-oriented wireless resources scheduling," *Journal of Network and Computer Applications*, vol. 158, p. 102594, 2020.

[18] A. A. Bisu, A. Purvis, K. Brigham, and H. Sun, "A framework for end-to-end latency measurements in a satellite network environment," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[19] J. Shi, H. Yang, C. Pan, X. Chen, Q. Sun, Z. Yang, and W. Xu, "Low-latency design for satellite assisted wireless vr networks," *IEEE Communications Letters*, vol. 27, no. 6, pp. 1555–1559, 2023.

[20] H. Nguyen-Kha, V. N. Ha, E. Lagunas, S. Chatzinotas, and J. Grotz, "Leo-to-user assignment and resource allocation for uplink transmit power minimization," 2023.

[21] L. Chen, V. N. Ha, E. Lagunas, L. Wu, S. Chatzinotas, and B. Ottersten, "The next generation of beam hopping satellite systems: Dynamic beam illumination with selective precoding," *IEEE Transactions on Wireless Communications*, 2022.

[22] V. N. Ha, Z. Abdullah, G. Eappen, J. C. M. Duncan, R. Palisetty, J. L. G. Rios, W. A. Martins, H.-F. Chou, J. A. Vasquez, L. M. Garces-Socarras, H. Chaker, and S. Chatzinotas, "Joint linear precoding and dft beamforming design for massive mimo satellite communication," in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1121–1126.

[23] Y. Guo, Q. Yang, F. Fu, and K. S. Kwak, "Quality-oriented rate control and resource allocation in dynamic ofdma networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.

[24] Y. Guo, Q. Yang, and K. S. Kwak, "Quality-oriented rate control and resource allocation in time-varying ofdma networks," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 2324–2338, 2017.

[25] M. Guerster, J. Grotz, P. Belobaba, E. Crawley, and B. Cameron, "Revenue management for communication satellite operators - opportunities and challenges," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–15.

[26] Y. Zhu, M. Sheng, J. Li, and R. Liu, "Performance analysis of intermittent satellite links with time-limited queuing model," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2282–2285, 2018.

[27] P. Schulz, "Queueing-theoretic end-to-end latency modeling of future wireless networks," 2020.

[28] P. Schulz, L. Ong, P. Littlewood, B. Abdullah, M. Simsek, and G. Fettweis, "End-to-end latency analysis in wireless networks with queuing models for general prioritized traffic," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2019, pp. 1–6.

[29] N. J. H. Marcano, L. Diez, R. A. Calvo, and R. H. Jacobsen, "On the queuing delay of time-varying channels in low earth orbit satellite constellations," *IEEE Access*, vol. 9, pp. 87 378–87 390, 2021.

[30] M. Irazabal, E. Lopez-Aguilera, I. Demirkol, and N. Nikaein, "Dynamic buffer sizing and pacing as enablers of 5g low-latency services," *IEEE Transactions on Mobile Computing*, vol. 21, no. 3, pp. 926–939, 2022.

[31] Y. Zhu, M. Sheng, J. Li, D. Zhou, and Z. Han, "Modeling and performance analysis for satellite data relay networks using two-dimensional markov-modulated process," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3894–3907, 2020.

[32] A. Anand and G. de Veciana, "Resource allocation and harq optimization for urllc traffic in 5g wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.

[33] W. Whitt, "Time-varying queues," *Queueing models and service management*, vol. 1, no. 2, 2018.

[34] M. Defraeye and I. Van Nieuwenhuyse, "Controlling excessive waiting times in small service systems with time-varying demand: An extension of the isa algorithm," *Decision Support Systems*, vol. 54, no. 4, pp. 1558–1567, 2013, rapid Modeling for Sustainability. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923612001790

[35] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt, "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[36] Z. Lin, Z. Ni, L. Kuang, C. Jiang, and Z. Huang, "Dynamic beam pattern and bandwidth allocation based on multi-agent deep reinforcement learning for beam hopping satellite systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 3917–3930, 2022.

[37] E. S. Furman, "Models for capacity allocation in anticipation of time-varying demand," 2020.

[38] D. Armbruster, S. Göttlich, and S. Knapp, "Continuous approximation of $m\_t/m\_t/1$ distributions with application to production," *arXiv preprint arXiv:1807.07115*, 2018.

[39] B. Han, V. Sciancalepore, X. Costa-Perez, D. Feng, and H. D. Schotten, "Multiservice-based network slicing orchestration with impatient tenants," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 5010–5024, 2020.

[40] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for urllc and embb coexistence in multi-user 5g networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[41] A. Benjebbour, K. Kitao, Y. Kakishima, and C. Na, "3gpp defined 5g requirements and evaluation conditions," *NTT DOCOMO Technical Journal*, vol. 19, no. 3, pp. 13–23, 2018.

[42] V. N. Ha, T. T. Nguyen, L. B. Le, and J.-F. Frigon, "Admission control and network slicing for multi-numerology 5g wireless networks," *IEEE Networking Letters*, vol. 2, no. 1, pp. 5–9, 2020.

[43] G. F. Newell, "Queues with time-dependent arrival rates i—the transition through saturation," *Journal of Applied Probability*, vol. 5, no. 2, pp. 436–451, 1968.

[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787

[45] H. Hindi, "A tutorial on convex optimization ii: duality and interior point methods," in *2006 American Control Conference*, 2006, pp. 11 pp.–.

[46] T. T. Nguyen, V. N. Ha, L. B. Le, and R. Schober, "Joint data compression and computation offloading in hierarchical fog-cloud systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 293–309, 2020.

[47] X. Xu, Q. Wang, C. Liu, and C. Fan, "A satellite network data transmission algorithm based on adaptive lt code," in *2021 International Conference on Space-Air-Ground Computing (SAGC)*, 2021, pp. 100–105.

[48] U. Speidel and L. Qian, "Striking a balance between bufferbloat and tcp queue oscillation in satellite input buffers," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[49] N. Torkzaban, A. Zoulkarni, A. Gholami, and J. S. Baras, "Capacitated beam placement for multi-beam non-geostationary satellite systems," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[50] V. K. Gupta, V. N. Ha, E. Lagunas, H. Al-Hraishawi, L. Chen, and S. Chatzinotas, "Combining time-flexible geo satellite payload with precoding: The cluster hopping approach," *IEEE Transactions on Vehicular Technology*, pp. 1–15, 2023.

[51] V. N. Ha, T. T. Nguyen, E. Lagunas, J. C. Merlano Duncan, and S. Chatzinotas, "GEO payload power minimization: Joint precoding and beam hopping design," in *GLOBECOM 2022 - 2022 IEEE Global Commun. Conf.*, 2022, pp. 6445–6450.

[52] McKinsey, "0.1 cents per MB: Ensuring future data profitability in emerging markets," https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/telecoms/pdfs/recall_no17_cost_per_mb.ashx, 2023, [Online; accessed 13-July-2023].

[53] M. Mozaffari, Y.-P. E. Wang, and K. Kittichokechai, "Blocking probability analysis for 5g new radio (nr) physical downlink control channel," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

[54] ITU-R, "Detailed specifications of the radio interfaces for the satellite component of the International Mobile Telecommunications-2000 (IMT-2000) ," https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.1850-0-201001-S!!PDF-E.pdf, 2010.

[55] W. A. Massey, "The analysis of queues with time-varying rates for telecommunication models," *Telecommunication Systems*, vol. 21, pp. 173–204, 2002.

[56] O. B. Jennings and W. A. Massey, "A modified offered load approximation for nonstationary circuit switched networks," *Telecommunication Systems*, vol. 7, pp. 229–251, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:35513460

[57] H. Abu-Ghazaleh and A. S. Alfa, "Channel assignments in wireless networks with time-varying traffic behaviors," in *2015 8th IFIP Wireless and Mobile Networking Conference (WMNC)*, 2015, pp. 285–292.

[58] L. Duan, J. Huang, and B. Shou, "Duopoly competition in dynamic spectrum leasing and pricing," *IEEE Transactions on Mobile Computing*, vol. 11, no. 11, pp. 1706–1719, 2012.

[59] K. Wang, F. Fang, D. B. d. Costa, and Z. Ding, "Sub-channel scheduling, task assignment, and power allocation for oma-based and noma-based mec systems," *IEEE Transactions on Communications*, vol. 69, no. 4, pp. 2692–2708, 2021.

[60] F. Kavehmadavani, V.-D. Nguyen, T. X. Vu, and S. Chatzinotas, "Intelligent traffic steering in beyond 5g open ran based on lstm traffic prediction," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[61] M. Camelo, P. Soto, and S. Latré, "A general approach for traffic classification in wireless networks using deep learning," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 5044–5063, 2022.

[62] A. Abada, B. Yang, and T. Taleb, "Traffic flow modeling for uav-enabled wireless networks," in *2020 International Conference on Networking and Network Applications (NaNA)*, 2020, pp. 59–64.

Teweldebrhan Mezgebo Kebedew (Student Member, IEEE) received the MSc. degree in Telecommunication Network Engineering from Addis Ababa University, Ethiopia, in 2020. Presently, he is pursuing a Ph.D. degree with the Interdisciplinary Center for Security, Reliability, and Trust at the University of Luxembourg, Luxembourg. He served as a Radio Access Network (RAN) Optimization Engineer at Ethio Telecom, Ethiopia. His current research interests include resource allocation in 5G/B5G satellite communication (satcom) networks and the application of artificial intelligence for optimizing resources.

Vu Nguyen Ha (Member, IEEE) received the B.Eng. degree (Hons.) from the French Training Program for Excellent Engineers in Vietnam, Ho Chi Minh City University of Technology, Vietnam, the Addendum degree from the École Nationale Supérieure des Télécommunications de BretagneGroupe des École des Télécommunications, Bretagne, France, in 2007, and the Ph.D. degree (Hons.) from the Institut National de la Recherche Scientifique-Énergie, Matériaux et Télécommunications, Université du Québec, Montreal, QC, Canada, in 2017. From 2016 to 2021, he worked as a Postdoctoral Fellow with the Ecole Polytechnique de Montreal, and then the Resilient Machine Learning Institute, École de Technologie Supérieure, University of Québec. He is currently a Research Scientist with the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg. His research interests include applying/developing optimization and machine-learning-based solution for RRM problems in MAC/PHY layers of several wireless communication systems, including SATCOM, 5G/beyond-5G, HetNets, Cloud RAN, massive MIMO, mobileedge computing, and 802.11ax WiFi. He received the Innovation Award for his Ph.D. degree. He was a recipient of the FRQNT Postdoctoral Fellowship for International Researcher (PBEEE) awarded by the Québec Ministry of Education, Canada, in 2018 and 2019. In 2021 and 2022, he was also awarded the Certificate for Exemplary Reviews by the IEEE WIRELESS COMMUNICATIONS LETTERS.

Eva Lagunas (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2010 and 2014, respectively. She has held positions at UPC, Centre Tecnologic de Telecomunicacions de Catalunya (CTTC), University of Pisa, Italy; and the Center for Advanced Communications (CAC), Villanova University, PA, USA. In 2014, she joined the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, where she currently holds a Research Scientist position. Her research interests include terrestrial and satellite system optimization, spectrum sharing, resource management and machine learning.

Joël Grotz (Senior Member, IEEE)) received the degree in electrical engineering from the University of Karlsruhe and the Grenoble Institute of Technology in 1999, and the Ph.D. degree in telecommunications jointly from the University of Luxembourg and KTH, Stockholm, in 2008. He worked with SES, Betzdorf, Luxembourg, on the development of satellite broadband communication system design for GEO and MEO high-throughput satellite systems on different ground segment and space segment topics and system optimization aspects. He worked with the Technical Labs at ST Engineering iDirect (former Newtec Cy at Sint-Niklaas in Belgium) on topics of system design and signal processing in satellite modems. He is currently working as a Senior Manager at SES on the development of a dynamic resource management system for novel flexible satellite systems, including SES-17 and O3b mPOWER as well as future satellite systems under planning.

Symeon Chatzinotas (Fellow, IEEE) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, U.K., in 2006 and 2009, respectively. He is currently a Full Professor/Chief Scientist I and the Head of the Research Group SIGCOM, Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg. In parallel, he is an Adjunct Professor with the Department of Electronic Systems, Norwegian University of Science and Technology and a Collaborating Scholar with the Institute of Informatics and Telecommunications, National Center for Scientific Research “Demokritos.” In the past, he has lectured as a Visiting Professor with the University of Parma, Italy and contributed in numerous research and development projects for the Institute of Telematics and Informatics, Center of Research and Technology Hellas and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has authored more than 700 technical papers in refereed international journals, conferences, and scientific books and has received numerous awards and recognitions, including the IEEE Fellowship and an IEEE Distinguished Contributions Award. He is currently on the editorial board of the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and the International Journal of Satellite Communications and Networking.