A Data-Driven Assessment Framework for QoS Booking in Real 5G Networks

David Reiss
Signal Theory and Communications
Universitat Politècnica de Catalunya
Barcelona, Spain
david.reiss@upc.edu

Oriol Sallent
Universitat Politècnica de Catalunya
i2CAT Foundation
Barcelona, Spain
jose.oriol.sallent@upc.edu

Miguel Catalan-Cid
Mobile Wireless Internet Group
i2CAT Foundation
Barcelona, Spain
miguel.catalan@i2cat.net

Daniel Camps-Mur
Mobile Wireless Internet Group
i2CAT Foundation
Barcelona, Spain
daniel.camps@i2cat.net

***Abstract*—The transition toward the Network-as-a-Service (NaaS) paradigm enables programmable telco capabilities via standardized northbound interfaces, such as the CAMARA QoS Booking or the slicing booking APIs. However, exposing advance resource reservation capabilities to third-party applications introduces substantial challenges to Radio Access Network (RAN) stability due to the inherent uncertainty in future traffic profiles, user mobility, and localized radio conditions. This paper proposes a comprehensive, data-driven stochastic framework for assessing the feasibility of QoS booking requests. The proposed assessment algorithm characterizes admission feasibility via a dual-metric approach that captures both the frequency and severity of capacity excess while also contemplating the preemption of radio resources allocated to best-effort devices to maximize network utilization and revenue. Leveraging an eight-week dataset of cell-level Key Performance Measurements (KPMs) from a commercial 5G New Radio (NR) n78 band deployment, the framework constructs empirical probability distributions of background traffic and Channel Quality Indicator (CQI)/MIMO Rank profiles to derive the aggregate utilization distribution at booking execution time. The evaluation across two distinct urban deployment scenarios demonstrates that a cell's effective spectral efficiency dominates booking feasibility over raw background utilization. In addition, results reveal that the proposed flexible preemption mechanism safely scales concurrent booking acceptance, doubling it for scenarios with good channel conditions, while minimizing resource under-utilization.**

***Index Terms*—Network-as-a-Service (NaaS), CAMARA APIs, QoS Booking, Radio Resource Management, 5G RAN**

I. Introduction

The Sixth Generation (6G) of mobile communication networks is expected to deliver substantial capacity enhancements, significantly broadening overall network capabilities. In order to fully exploit such capabilities, the telco industry is increasingly focusing on promoting network innovation and programmability. In this context, the Network as a Service (NaaS) paradigm rises as a prominent approach to expose low-level network functionalities as programmable services that can be dynamically consumed by third-parties.

The GSMA Open Gateway initiative has emerged as the telecom industry's main collaborative effort to accelerate NaaS adoption, bringing together mobile network operators (MNOs), technology vendors, and developer communities [1]. Within the Open Gateway ecosystem, one of the main enablers is the CAMARA project [2]. CAMARA is an open-source initiative from the Linux Foundation which aims at standardizing and exposing telco network capabilities via open Northbound Application Programming Interfaces (APIs) that simplify access to Core and Radio Access Network (RAN) functionalities

By abstracting the complexity of underlying network systems, CAMARA defines Service and Service Management APIs to enable seamless application integration and consistent operation across heterogeneous network environments. Organized into multiple API families, including communication quality, fraud prevention and authentication, location services, and computational services, and closely aligned with the GSMA Operator Platform API Group (OPAG) [3], CAMARA provides a suitable framework for exposing network capabilities in B5G and 6G environments.

While fraud prevention and authentication APIs have attracted significant industry interest in recent years, mainly due to their relevance for third parties, such as banks, and their comparatively lower implementation complexity, recent market studies indicate growing momentum around communication quality APIs [4]. This trend reflects increasing interest from enterprises and developers in exploiting the differentiated connectivity capabilities enabled by 5G, and is consistent with industry reports showing that network-slicing-based differentiated connectivity offerings are entering the market [5]. However, unlike information-exposure APIs, communication quality APIs may require active network control to provide guaranteed service levels, including slice

This work is supported by the Grant TRAINER-6G (PID2023-146748OB-I00) funded by MCIN/AEI/10.13039/501100011033, and by the European Union through the SOVEREIGN-6G project (GA No. 101292787).

979-8-3195-0167-7/26/$31.00 ©2026 IEEE

configuration, resource reservation, admission control, and SLA monitoring. Their adoption therefore introduces significant RAN management challenges, since guaranteed connectivity for selected services must be provided without degrading resource efficiency or the performance experienced by other users.

Within this context, QoS Booking represents a particularly relevant communication quality use case, as it allows applications to request guaranteed network performance for a given location, time window, number of devices, and QoS profile. Unlike QoS on-demand (QoD) enforcement, QoS Booking requires the operator to assess resource availability before the service is activated, considering both the expected network state during the requested interval and the impact of the additional load introduced by the booked devices. This shifts the problem from real-time QoS configuration to advance feasibility assessment under traffic and radio uncertainty.

Existing approaches mainly focus on QoD API execution, slice reservation and admission control, or predictive capacity assessment based on deterministic forecasts. To the best of our knowledge, no prior work has specifically addressed the operator-side feasibility assessment of CAMARA QoS Booking requests in live commercial RANs. In this scenario, the operator must decide in advance whether a service request can be safely admitted under probabilistic background-load and radio-efficiency conditions, while explicitly controlling the impact on best-effort traffic.

In this paper, we propose an assessment framework to support the QoS booking API, defining a systematic procedure to assess the feasibility of requests submitted by end API consumers. Specifically, the main contributions of this work are threefold:

- We propose a data-driven algorithm to evaluate whether an incoming QoS booking request should be accepted or rejected. The framework assesses admission viability by balancing both the statistical probability of a capacity excess and the magnitude of such excess.
- We include within the assessment framework a preemption factor to reallocate resources from best-effort background traffic to the incoming QoS booking requests. This allows operators to safely maximize network utilization and revenue while managing potential service degradation.
- We validate the proposed assessment framework utilizing real-world network telemetry from a cellular deployment provided by a major Spanish Mobile Network Operator (MNO).

Overall, this framework provides a robust mechanism to balance efficient radio resource management while supporting the reliable provisioning of QoS-enabled services.

The rest of the paper is organized as follows. Section II presents the related work and positions the proposed feasibility assessment within the CAMARA framework. Then, Section III presents the QoS booking request format and details the mathematical description of the assessment framework. After that, Section IV provide the evaluation of the proposed assessment methodology utilizing real cellular data. Finally, Section V presents the conclusions and future work.

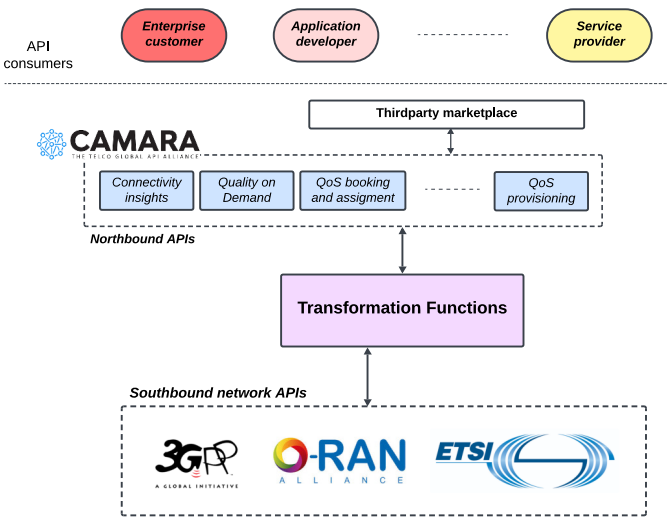


Fig. 1: Feasibility assessment positioning within CAMARA framework.

II. Related Work

The CAMARA project abstracts underlying network complexities to deliver demand-driven NaaS APIs to enterprise and application developers. These capabilities are exposed via a northbound ecosystem marketplace, allowing consumers to efficiently discover and request specific telco operator capabilities.

Central to this architecture is the Transformation Function (TF), a critical logic layer situated beneath the Service and Service Management APIs. The TF serves as the deterministic link between northbound and southbound interfaces, responsible for mapping high-level developer-centric requests into low-level calls for 3GPP or ETSI APIs, such as the 3GPP Network Exposure Function (NEF) [6], or the ETSI MEC APIs [7]. For example, CAMARA Quality on Demand (QoD) requests can be translated by the Transformation Function into NEF calls such as *AsSessionWithQoS*. These calls initiate a Policy Control Function (PCF) procedure to configure the required QoS treatment toward the RAN and the target UEs, including parameters such as flow priority and maximum bitrate [2].

Figure 1 illustrates the high-level CAMARA workflow. While CAMARA defines the northbound APIs and the functional requirements for TFs, their southbound communication with lower layers such as the 5G core follows standardized APIs and data model. However, the specific implementation of TFs remains undefined, allowing network operators to deploy proprietary logic tailored to their internal infrastructure and to their exposure and control capabilities. This implementation flexibility motivates the proposed assessment framework.

Existing research on CAMARA-based exposure has so far focused mainly on the integration of vertical use cases within standardized network architectures, particularly on the implementation of TFs that connect northbound CAMARA APIs with southbound 3GPP or ETSI interfaces. For instance,

authors in [8] present a 3GPP-compliant architecture and implementation of CAMARA Device Location and QoD APIs to enable dynamic QoS adaptation for Connected and Automated Mobility (CAM) scenarios. They integrate these exposed APIs with a V2X Application Server at MEC and demonstrate in a testbed that high-priority QoS profiles improve selected uplink traffic flows. Similarly, authors in [9] propose a mechanism that utilizes QoD API integrated within an O-RAN compliant testbed. The proposed solution enables video streaming UEs to request temporary bandwidth boosts to improve Quality of Experience (QoE) under highly congested scenarios.

Although these works provide representative examples of how CAMARA logic can be integrated with current network architectures, they mainly address the execution of QoD-related API requests under controlled or testbed conditions, demonstrating the feasibility and benefits of CAMARA workflows and APIs. However, applying QoS guarantees in realistic environments requires operators to assess whether the required resources are available, considering both the impact of the current network state on the requested flows and the impact of the new flows on existing services. This becomes particularly relevant in QoS booking scenarios, where operators must decide in advance whether future requests can be safely admitted under uncertain traffic demand, radio conditions, and resource availability. This motivates feasibility assessment mechanisms that account for the variable behavior of live commercial RANs, rather than only defining the API translation and enforcement logic.

Related problems have also been studied in the context of network slicing admission control, advance reservation, and predictive resource provisioning. A prioritized admission control and resource reservation framework is proposed in [10], where slice requests can be processed before their activation time while accounting for uncertain user demand, time-varying resource requirements, and the impact of reservations on background services. Advance slice reservation is addressed from a scheduling perspective in [11], where requests are characterized by QoS requirements, duration, and timing constraints, and the orchestrator decides where and when to admit them. A two-timescale slicing framework is proposed in [12], separating long-timescale resource reservation from short-timescale RAN allocation while explicitly considering traffic variation and channel uncertainty. Closer to RAN dimensioning, [13] exploits per-gNodeB radio statistics, particularly Modulation and Coding Scheme (MCS) distributions, together with expected vehicular traffic to proactively reserve RAN resources for URLLC slices. More recently, a QoS and capacity prediction framework is proposed in [14], combining live-network measurements, interference forecasting, spectral-efficiency estimation, and Physical Resource Block (PRB) utilization prediction to support proactive admission of 5G slicing services.

These works show that admitting future services with performance guarantees requires accounting for demand, radio conditions, and resource availability before execution. However, they primarily address network-slicing reservation and provisioning problems, rather than the operator-side feasibility assessment. The closest work in the literature is [14], which predicts the PRB demand and cell load associated with future network slicing requests. However, its admission control decisions rely entirely on deterministic forecasts. These deterministic models fail to capture statistical diversity, leading to significant forecasting errors, particularly during peak rush hours, when traffic is more volatile. In contrast, this paper introduces a stochastic assessment framework that leverages full historical distributions derived from live commercial RAN measurements. This allows us to characterize both the probability and expected severity of PRB capacity excess under controlled best-effort traffic preemption. Consequently, this stochastic approach provides robust protection against misestimates in resource utilization and spectral efficiency. Based on this gap, the following section formalizes the QoS Booking use case and introduces the proposed feasibility assessment framework.

III. QoS booking Feasibility assessment framework

In this section, we present the core design of the feasibility assessment framework. We begin by providing a detailed description of the CAMARA QoS booking API request format, identifying the key attributes required for feasibility evaluation. Building on this interface specification, the subsequent subsection details the formal mathematical description of the assessment algorithm, outlining how these inputs are combined to model capacity constraints and resource preemption.

A. QoS booking request format

The motivation behind the CAMARA QoS booking and assignment API is to allow third-parties to request network performance guarantees in advance, enabling differentiated connectivity over a specified time interval. Among other, the target use cases of this interface include live event streaming, augmented/virtual reality experiences, autonomous vehicles support, and VIP or business communication services. Although this API is in an early-development stage, the expected inputs are already defined in the specification of the API [15].

First, the CAMARA consumer needs to consider the QoS profile that most adjusts to its own service requirements. This is done through the CAMARA QoS profiles API [16], which presents to the consumer the different profiles offered by the API providers, namely the MNOs. The main parameters of these profiles include uplink (UL) and downlink (DL) required data rates, jitter, packet-delay budget, and priority level, among others, becoming high-level representations of the 5G System QoS flow model defined by 3GPP [17]. Throughout this paper, we focus on QoS bookings targeting the DL throughput, although generalizing the proposed framework to the UL traffic is straightforward.

Then, the inputs to the QoS booking API include the flow identification parameters (e.g., IP of the device, application server and application ports) the location of the booking, the time window (as date, start time and duration), the number of devices within the booking, and the selected QoS profile

TABLE I: Example of QoS Booking Request Parameters

Input Name	Units	Example
Location		
latitude	degrees	41.3851
longitude	degrees	2.1734
area type	string	"circle"
radius	meters	500
Start Time	string	"dd/mm/yyyy" "14:30:00"
Duration	seconds	3600
Devices	integer	12
QoS Profile		
minimum-DownlinkBitRate	kbps	5000

these devices will require. Table I illustrates an example of the required parameters contemplated in our analysis. In addition, the API documentation also provides the expected responses for any arbitrary request: i) successful, with the booking accepted synchronously, ii) failure, with the booking rejected synchronously, and iii) pending, if the response cannot be determined directly.

B. Feasibility assessment framework

In order to decide the acceptance or rejection of an incoming booking request days or weeks in advance for a given cell within the affected area we must consider: (i) the background PRB occupancy, $\eta_B = \eta_{B_{GBR}} + \eta_{B_{nGBR}}$, representing regular GBR and non-GBR traffic of devices within the same area as the requested one, (ii) the estimated increase in PRB occupancy to allocate the candidate booking request, $\Delta\eta$, and (iii) the estimated PRB occupancy increase due to previously accepted booking requests overlapping in space and time with the incoming one, $\Delta\eta^*$.

For the background PRB occupancy, η_B, we adopt a data-driven methodology on which historical Key Performance Measurement (KPM) counters are used to derive the empirical distribution of the PRB occupancy.

In contrast, $\Delta\eta$ must be computed according to the incoming request. As detailed above, in this paper we focus on QoS bookings that demand a guaranteed DL target data rate, T_r, for a certain number of devices N_{dev}. Thus, the PRB occupancy increment is estimated as follows:

$$\Delta\eta = \frac{T_r \cdot N_{dev}}{C_{PRB}} \tag{1}$$

where the C_{PRB} term represents the capacity per PRB during the requested time period. In particular, the capacity per PRB relies on the spectral efficiency in the target area during the requested time window. Based on [18] we compute C_{PRB} as:

$$C_{PRB} = \nu \cdot SE \cdot 11.2 \cdot \Delta f \cdot (1 - OH) \cdot DL_{split} \cdot 10^{-3} \tag{2}$$

where ν represents the number of MIMO layers, SE the spectral efficiency, Δf the sub-carrier spacing (SCS) in kHz, OH the overhead defined by 3GPP, and DL_{split} the ratio between UL/DL slots (if TDD is used). Note that, rather than being deterministic, the effective spectral efficiency term $\nu \cdot SE$ is a stochastic variable, depending on the channel conditions of a certain area and time period. Similarly as in [19], we also advocate for a data-driven approach on which both ν and SE are derived through real CQI and Rank (MIMO layers) cellular counters. Note that the $\Delta\eta^*$ distribution is computed using the exact same method.

For both η_B and the $\nu \cdot SE$, the framework adopts a profiling methodology in which empirical distributions are derived from historical telemetry matching the exact same day of the week, d, and hour window, h, as the requested booking period. This filtering captures recurrent temporal patterns in background utilization and radio-efficiency conditions that are specific to the target interval.

In addition, assuming whether the booking devices exhibit correlated mobility or not is crucial to compute $\Delta\eta$. In this work, we assume each device has independent channel conditions, meaning that $\Delta\eta = \sum_i^{N_{dev}} \Delta\eta_i$ rather than $N_{dev} \cdot \Delta\eta_i$, which has a strong impact on the variance and tail behavior of the resulting stochastic distribution of $\Delta\eta$. Note that, correlated devices are highly susceptible to simultaneous deep fades if the booking device group moves into poor coverage together. Conversely, uncorrelated devices benefit from statistical diversity, significantly reducing the joint probability of all devices experiencing poor channel conditions simultaneously, resulting in a narrower $\Delta\eta$ distribution. Since the API lacks this context, this assumption is arbitrary and should adapt to each booking request, highlighting a clear requirement for future versions of the QoS booking API.

Based on the previous considerations, the assessment framework operates on the aggregate utilization, $Z = \eta_{B_{nGBR}} + \eta_{B_{GBR}} + \Delta\eta^* + \Delta\eta$, which is modeled as a stochastic variable. Assuming variable independence, the distribution of Z is obtained as:

$$f_Z(z) = (f_{\eta_{nGBR}} * f_{\eta_{\text{GBR}}} * f_{\Delta\eta^*} * f_{\Delta\eta})(z) \tag{3}$$

The objective of the proposed framework is to determine whether a booking request should be accepted or rejected. To this end, the framework evaluates the probability that the aggregate utilization exceeds the available network capacity within the requested target area and during the specified booking period.

In addition, we introduce a preemption factor, $\alpha \in [0,1]$, applied to the non-GBR background traffic. We envision that, in order to maximize resource utilization and revenue, operators may choose to accommodate QoS bookings by preempting resources originally allocated to best-effort background traffic that can be offloaded to neighboring or coverage cells within the area.

Consequently, the assessment framework evaluates the capacity outage or the probability of an excess, ξ, under a given preemption factor as:

$$P(\xi(\alpha)) = P(\alpha \cdot \eta_{B_{nGBR}} + \eta_{B_{GBR}} + \Delta\eta^* + \Delta\eta \geq N_{PRB}), \tag{4}$$

where N_{PRB} represents the total DL PRBs. While the excess probability formulation tracks the frequency of capacity outages, it does not consider the severity of such. For instance, if the

aggregate utilization results in a narrow distribution centered at 100% capacity, the excess probability can report a large value while the expected magnitude of the capacity outage remains relatively low. To this end, we define a second condition with respect to the expected PRB excess, which is defined as follows:

$$\mathbb{E}[\xi(\alpha)] = \mathbb{E}[(Z - N_{PRB})^+] = \int_{N_{PRB}}^{\infty} (z - N_{PRB}) \cdot f_Z(z) dz \quad (5)$$

With all the above, the QoS booking request is accepted provided that (4) and (5) do not surpass certain thresholds. Moreover, the MNO policy may impose a lower bound on the amount of non-GBR traffic that must be preserved, denoted by α_{min}. This constraint ensures that best-effort users retain a minimum level of service, for instance in terms of achieved bitrate, even when resources are partially displaced to accommodate QoS bookings. Therefore, the objective of the assessment algorithm is to determine whether a feasible $\alpha^* \in [\alpha_{min}, 1]$ exists such:

$$\alpha^* = \max \left\{ \alpha \in [\alpha_{\min}, 1.0] \;\middle|\; P(\xi(\alpha)) \leq \theta_P,\, \mathbb{E}[\xi(\alpha)] \leq \theta_\xi \right\}, \quad (6)$$

where θ_P and θ_ξ represent the excess probability limit (frequency of the excess) and the expected PRB excess limit (magnitude of the excess), respectively.

Algorithm 1 describes the complete operation of the proposed assessment framework. It is composed of three main steps. First, we filter historical data and derive the empirical distributions of the background PRB occupancy and the joint CQI/Rank for the effective spectral efficiency, foundational to derive the C_{PRB}, and therefore the $\Delta\eta$ distribution.

Then, for each $\alpha \in [\alpha_{min}, 1]$ we scale the η_B distribution and derive $f_Z(z)$. Note that, in order to avoid the analytical complexity and potential numerical inaccuracies associated with resolving a convolution over sparse, non-uniform raw data arrays extracted from a cellular network dataset, we propose a Monte Carlo simulation consisting of K independent iterations, where the framework draws a single independent realization from the background traffic $\alpha \cdot \eta_B$ and the increment $\Delta\eta$ distributions.

Finally, when a feasible α holds the excess condition in (6), the algorithm returns an acceptance and the optimal parameter α^*, which quantifies the expected preemption factor at the booking execution time. Otherwise, a rejection occurs. Notably, when $\alpha^* = 1$ cells withing the target area have enough capacity to accommodate incoming requests without penalizing non-GBR background traffic.

The optimal preemption factor represents the maximum feasible fraction of non-GBR background traffic that can be preserved while accommodating the booking. Operators can utilize this outcome to derive the resource constraints required to accommodate accepted booking requests, such as network slicing configurations, minimizing the penalty imposed on best-effort traffic while maximizing overall resource efficiency. Note that, in cases on which α^* remains much higher than α_{min}, this estimate becomes crucial for operator's resource management planning.

Algorithm 1 End-to-End Assessment Algorithm

Input: d (Day), h (Hour), N_{dev} (Number of Devices), T_r (GBR demand)
Output: Decision (Accept/Reject), α^* (Expected Preemption Factor)

Initialize: $\alpha \leftarrow 1.0$
▷ *Step 1: Data Filtering & Distributions*
  Constrain historical data based on d and h
  Construct CQI/Rank distribution
  Build $\Delta\eta$ distribution given N_{dev} and T_r
while $\alpha \geq \alpha_{\min}$ **do**
▷ *Step 2: Monte Carlo Simulation*
    Build $\alpha \cdot \eta_{B_{nGBR}}$ scaled distribution
    Compute aggregate utilization distribution $f_Z(z)$ using a Monte Carlo simulation with K iterations
▷ *Step 3: Condition Evaluation*
    if $P(\xi(\alpha)) \leq \theta_P$ and $\mathbb{E}[\xi(\alpha)] \leq \theta_\xi$ **then**
      $\alpha^* \leftarrow \alpha$
      return (`ACCEPT`, α^*) **; EXIT**
    end if
    $\alpha \leftarrow \alpha - \Delta\alpha$
end while
return (`REJECT`, $\emptyset$) ▷ No feasible α

IV. QoS Booking Feasibility Evaluation

This section characterizes the operator's cellular dataset, detailing the two distinct deployment scenarios analyzed in this paper. Then, we evaluate various incoming booking requests across multiple capacity excess thresholds and preemption policies. Finally, we examine a multiple-booking overlap scenario to determine the maximum number of simultaneously feasible small booking requests.

A. Cellular network dataset: radio scenarios characterization

The dataset utilized in this paper, provided by a Spanish MNO, comprises cell-level KPMs reported with a fifteen-minute granularity during a two-month collection campaign in a Spanish city. In this analysis we leverage data from the n78 band carriers (3500 MHz), for which a total of 70 cells are included in the dataset. While data for lower-frequency carriers is available (e.g., 700 MHz, 2100 MHz), these are primarily utilized as coverage cells or not fully deployed, and are thus excluded from this analysis. We focus exclusively on the capacity layer, where QoS booking traffic is assumed to be accommodated.

TABLE II: Radio Configuration of the Analyzed Deployment

Parameter	Value / Configuration
Total Bandwidth	100 MHz
Available PRBs per cell	273 PRBs
Sub-Carrier Spacing, Δf	30 kHz
Operating Band	n78 (3.3 - 3.8 GHz)
Duplexing Scheme	TDD with 74% for DL, $N_{PRB} = 202$

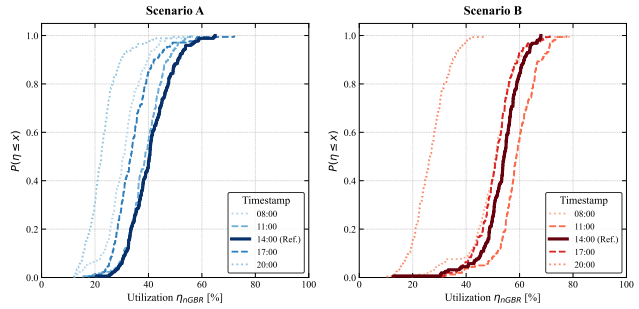


Fig. 2: Scenarios historical PRB utilization.

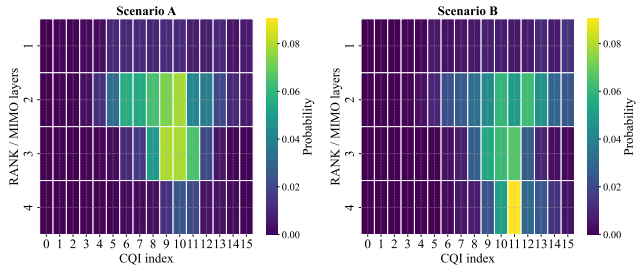


Fig. 3: Scenarios historical CQI/RANK distribution for Wednesday 11 - 13 hour range.

The radio specifications for the n78 band are illustrated in Table II. Cells belonging to this band have a total bandwidth of 100 MHz, resulting in 273 PRBs. However, they are configured using Time Division Duplexing (TDD) with a DDDSU pattern. This makes the DL split to represent 74% of the total PRBs.

From the full 70-cell dataset, the analysis presented in this paper focuses on two distinct scenarios (each composed of one individual cell), with strong differences in background PRB utilization, and spectral efficiency. Figure 2 presents the Cumulative Distribution Function (CDF) of historical PRB utilization in both scenarios, revealing significant disparities. While Scenario A maintains a lower utilization profile, with the probability of exceeding 50% utilization at 14 pm being around 0.1, Scenario B reports a higher utilization profile, with the same probability reaching 0.8 for the same hour. Furthermore, Scenario B exhibits greater temporal dispersion, with substantial utilization fluctuations observed between off-peak hours (20:00) and peak periods (11:00). Note that, the DL PRB utilization KPM allows us to derive the η_B distribution.

In addition, Figure 3 illustrates the 2D heatmap of the reported CQI indices and MIMO layer configurations for Wednesday 11 - 13 hour range (later used as the example booking period). We observe that Scenario A typically exhibits poor channel quality conditions when compared to Scenario B. In particular, Scenario A distribution peaks at CQI index 9-10 with 3 DL MIMO layers configured, while Scenario B peaks at CQI index 11 with 4 DL MIMO layers configured. This results in Scenario B having higher effective spectral efficiency, and, consequently, higher capacity per PRB, which is critical to accommodate incoming QoS bookings.

To map these distributions into the effective spectral efficiency, $\nu \cdot SE$, used in eq. 2, each CQI index is translated into its corresponding spectral-efficiency value according to Table 2 of the 3GPP specification [20], which matches the reporting configuration used by the devices in the n78 band of the dataset. Together with the radio parameters described in Table II, this mapping yields the capacity per PRB distribution over the requested time period, and consequently, the $\Delta\eta$ distribution.

Note that, scenarios combining low spectral efficiency with high background load would generally lead to systematic rejection for demanding bookings, whereas scenarios with high spectral efficiency and low background load would generally result in straightforward acceptance. On the other hand, the selected cases represent non-trivial cases where background utilization and radio efficiency create competing effects.

To evaluate the framework under different demand conditions, we define five different QoS booking examples which are requested for a Wednesday from 11 am to 13 am. As defined in Table III, these examples range from relaxed to stringent, capacity demanding reservations, according to N_{dev} and T_r parameters.

Considering that GBR traffic still remains a low portion of the public network traffic, all the traffic reported in the dataset is considered as non-GBR, $\eta_{B_{GBR}} = 0$. Moreover, multi-cell QoS booking requests are also left outside the scope of this paper. Therefore, the analysis is conducted over each individual scenario, without considering inter-cell mobility of devices.

B. Initial feasibility assessment

First, we start by analyzing the results without considering QoS booking requests overlap and under a no-preemption policy, where $\alpha_{min} = 1$, and use a Monte Carlo simulation with $K = 10^5$ iterations. Figure 4 shows the Z distribution obtained for each scenario and booking example. The boxplot illustrates the 5th-to-95th percentile range, with the wider color box being the 25-75 percentile range. Typically, Scenario A exhibits a wider deviation than Scenario B, revealing that not only has poorer radio conditions but also less certain historical data. We observe that as the booking requests become more capacity-demanding, Scenario A also surpasses Scenario B in the expected aggregate utilization. This shift occurs because Z becomes predominantly driven by $\Delta\eta$ rather than by η_B. Moreover, bookings 1, 2 and 3 seem to have a very low excess probability and could be directly accepted. On contrary, bookings 4 and 5 present acceptance complexities in both scenarios if no preemption factor is allowed.

TABLE III: QoS booking examples and no-preemption excess analysis

QoS BookingID	N_{dev}	T_r (Mbps)	Scenario A		Scenario B	
			P_{excess}	$\xi(\%)$	P_{excess}	$\xi(\%)$
1	5	5	0.005	0.16%	0.002	0.09%
2	10	5	0.011	0.41%	0.005	0.21%
3	25	5	0.069	2.05%	0.053	0.89%
4	20	10	0.621	11.03%	0.716	8.01%
5	20	15	0.997	42.48%	0.999	30.78%

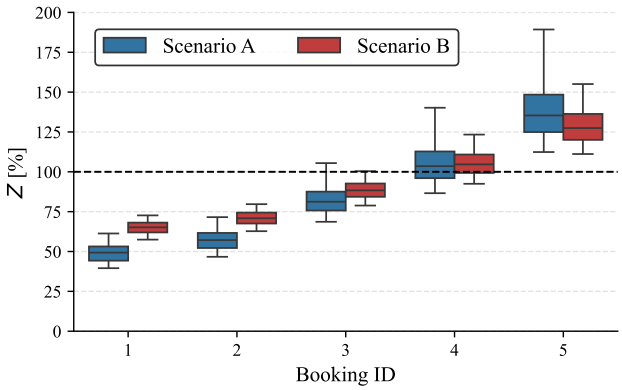


Fig. 4: Aggregate utilization of the booking requests with $K = 10^5$ iterations.

In this regard, Table III presents the detailed results for the defined $P(\xi)$ and $\mathbb{E}[\xi]$, extending the analysis to the full distribution rather than limiting it to the 5th-to-95th percentile range. First, we observe a huge disparity between the excess probability and the expected excess across all booking examples, highlighting the necessity of our dual-metric approach. Moreover, results confirm that bookings 1, 2, 3 are likely to be accepted. On contrary, for the capacity-demanding bookings 4 and 5, both the excess probability and the expected excess boost in both scenarios. This highlights that under a no-preemption policy these bookings would be likely rejected.

Note that, a rejection of bookings 4 and 5 would translate into radio resource under-utilization. For instance, if a reject occurs, the expected PRB utilization during the booking execution time will be the expected value of $\eta_{B_{nGBR}}$. Thus, the expected unused capacity $N_{PRB} - \mathbb{E}[\eta_{B_{nGBR}}]$ becomes 58% for Scenario A and 41% for Scenario B.

C. Preemption-Aware Feasibility Assessment

The previous results are computed under a full protection of non-GBR resources. However, the proposed assessment framework allows for the modeling of such preemption to reallocate best-effort traffic resources to incoming booking requests through α. Figure 5 illustrates $P(\xi)$ and $\mathbb{E}[\xi]$ for $\alpha \in [0, 1]$ for the booking requests 4 and 5 across both scenarios.

Because these two booking examples are the most capacity-demanding, Scenario A suffers from higher excess probability when compared to Scenario B. As the requested booking capacity increases, Z becomes strongly dependent on $\Delta\eta$. Consequently, despite having lower background PRB occupancy, scenarios with worse historical radio channel conditions (i.e., lower effective spectral efficiency) require a higher preemption to accept such requests.

For instance, we observe that increasing the preemption of best-effort resources significantly reduces the excess metrics, making both scenarios likely to accept booking 4. Specifically, Scenario A achieves $P(\xi) < 0.15$ and $\mathbb{E}[\xi] < 5\%$ at $\alpha = 0.5$, whereas Scenario B achieves the same results with an α value

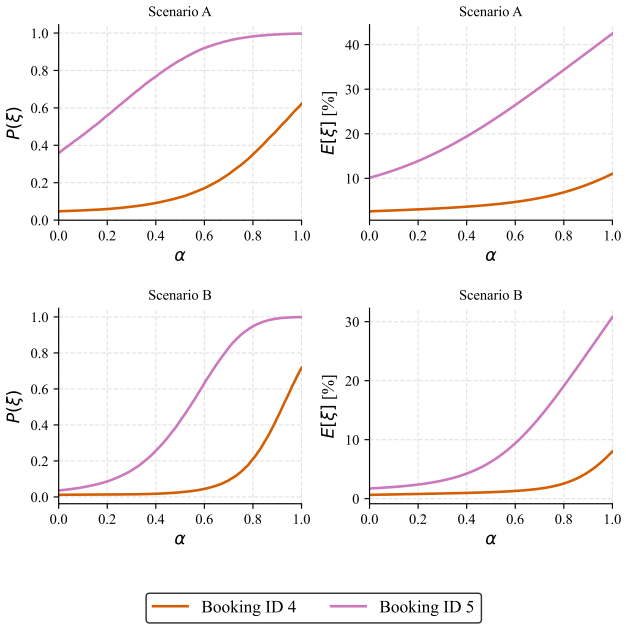


Fig. 5: Capacity demanding bookings excess analysis over α.

between 0.75 and 0.8. This highlights the inherent advantages of superior channel conditions.

On contrary, we observe that Scenario A is unable to reduce the excess conditions for booking 5. Even with full preemption of best-effort resources ($\alpha = 0$), the excess probability remains above 0.3, with an expected excess of 10%. On the other hand, Scenario B is able to sufficiently reduce the excess metrics under stringent α scaling. Nonetheless, depending on the minimum acceptable preemption limit ($\alpha_{\min}$) chosen by the operator, Scenario B might also trigger a rejection.

D. Multi-Booking Admission Capacity

The previous sections analyze the acceptance or rejection of simple booking requests. Now we analyze how many concurrent bookings can be accommodated in each scenario. Concretely, we utilize booking 1 as the base unit and evaluate the number of simultaneously accepted requests over multiple α, θ_P, θ_ξ.

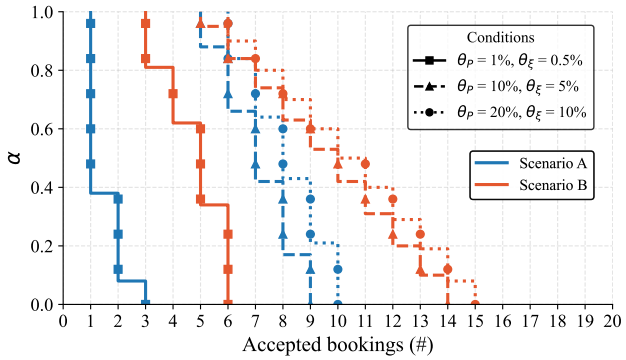


Fig. 6: Aggregate utilization of the booking requests.

Figure 6 provides a comprehensive overview of the two operational dimensions through which the operator can dynamically control the admission or rejection of incoming QoS booking requests. As demonstrated across all evaluated parameters, Scenario B consistently outperforms Scenario A. Specifically, even under the most conservative excess thresholds and highly restrictive preemption allowances, Scenario B maintains the capacity to accommodate three simultaneous bookings while Scenario A only one. Ultimately, these results reveal that a cell's effective spectral efficiency plays a more dominant role in determining QoS booking feasibility than its occupancy.

In addition, we observe the same trend as in Figure 5, where resource preemption yields a more pronounced impact on Scenario B. For instance, assuming an $\alpha_{\min} = 0.5$, Scenario B doubles the number of accepted bookings across all excess conditions, increasing from 5–6 to 10–11 accepted requests under relaxed thresholds. Conversely, applying the same logic to Scenario A leads to almost no improvement, merely resulting in two additional accepted bookings under the same relaxed excess thresholds. This highlights a crucial takeaway: although Scenario A is less utilized, consequently leading to a less resource efficient situation if a reject occurs, triggering preemption mechanisms again remains significantly more advantageous in cells possessing better channel conditions from the perspective of accepted GBR requests, i.e., increasing network revenue, and across stringent or relaxed excess constraints.

V. Conclusions and future work

In this paper we proposed a data-driven feasibility assessment framework to support the commercial adoption of the CAMARA QoS Booking API. The framework evaluates dual-metric excess conditions and incorporates a best-effort preemption factor to handle stochastic utilization and radio dynamics. Utilizing production telemetry from a live 5G NR deployment, our empirical evaluation of two distinct network scenarios yields several key operational insights. First, the results validate the necessity of our dual-metric assessment methodology, demonstrating that evaluating probability of excess without considering the magnitude of such fails to capture the true congestion profile. Second, our findings reveal that the effective spectral efficiency exerts a far more dominant role in determining QoS booking feasibility than the background PRB occupancy. Finally, the evaluation proves that introducing a flexible preemption factor can boost the acceptance of capacity demanding and concurrent QoS booking requests, therefore balancing best-effort penalty and cell resource utilization.

For the future work, we first aim to transition the best-effort preemption coefficient from a static term into a QoS-driven factor directly mapped from cellular counters. Considering historical data rate becomes crucial to derive an α_{min} linked to a minimum data rate for best-effort under individual cell conditions. Second, the assessment framework will be scaled to contemplate booking requests affecting a wider area, where multiple-cells will be involved. Improving in such direction requires the framework to consider inter-cell mobility patterns and a combined radio channel statistic, where cells with diverging channel conditions can be affected by the same request. Moreover, we pretend to extend also into evaluating booking overlaps of different requests types in terms of GBR and contextual conditions, where bookings can overlap only in a given portion of the full hour range and target area.

References

[1] GSMA, "The Ecosystem for Open Gateway NaaS API Development," *White paper*, 2023.

[2] J. Ordonez-Lucena and F. Dsouza, "Pathways towards network-as-a-service: the CAMARA project," in *Proceedings of the ACM SIGCOMM Workshop on Network-Application Integration*, 2022.

[3] GSMA, "OPG.02 – Operator Platform: Requirements and Architecture, Version 10.0," *Technical Specification*, 2025.

[4] Juniper Research, "Network API market, 2025-2023," *White Paper*, 2025.

[5] Ericsson, "Ericsson Mobility Report: June 2026," Ericsson, Tech. Rep., Jun. 2026, accessed: 2026-06-16. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2026

[6] 3rd Generation Partnership Project (3GPP), "TS 29.522: 5G System; Network Exposure Function Northbound APIs; Stage 3," 3GPP, Technical Specification (TS) Release 18, 2025.

[7] European Telecommunications Standards Institute (ETSI), "ETSI GS MEC 011: Multi-access Edge Computing (MEC); Edge Platform Application Enablement," ETSI, Group Specification (GS) V4.1.1, 2025.

[8] B. Turkovic, R. van de Vlasakker, S. Potnuru, Y. Shiferaw, R. S. Schwartz, P. Schackmann, G. Kovács, A. Forrai, G. Kakes, and P. Span, "Dynamic qos adaptation via exposed service apis for advanced cam use cases," in *2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2025.

[9] T. Tran, D. Gageot, C. Neumann, G. Bichot, A. Tlili, K. Boutiba, and A. Ksentini, "On the benefits and caveats of exploiting quality on demand network APIs for video streaming," in *NOSSDAV '24: Proceedings of the 34th edition of the Workshop on Network and Operating System Support for Digital Audio and Video*. ACM, 2024.

[10] Q.-T. Luu, S. Kerboeuf, and M. Kieffer, "Admission Control and Resource Reservation for Prioritized Slice Requests With Guaranteed SLA Under Uncertainties," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 3136–3153, 2022.

[11] R. Prasad and A. Sunny, "Scheduling Slice Requests in 5G Networks," *IEEE/ACM Transactions on Networking*, vol. 31, no. 6, pp. 3025–3036, 2023.

[12] H. Zhang and V. W. S. Wong, "A Two-Timescale Approach for Network Slicing in C-RAN," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6656–6669, 2020.

[13] N. Naddeh, S. Ben Jemaa, S. E. Elayoubi, and T. Chahed, "Anticipatory Slice Resource Reservation for 5G Vehicular URLLC Based on Radio Statistics," in *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2022, pp. 1019–1025.

[14] N. R. Moreno, F. Dsouza, A. Kassler, F. Pullem, B. Xu, M. Amend, and C. Choi, "QoS and Capacity Prediction for 5G Network Slicing," in *2025 21st International Conference on Network and Service Management (CNSM)*, 2025, pp. 1–5.

[15] CAMARA Project. (2025, Sep) Qos booking and assignment API specification, v0.1.0. Accessed: Mar. 18, 2026. [Online]. Available: https://github.com/camaraproject/QoSBooking

[16] CAMARA Project. (2025, Sep) Qos profiles API specification, v1.1.0. Accessed: Mar. 18, 2026. [Online]. Available: https://github.com/camaraproject/QualityOnDemand

[17] 3GPP, "System architecture for the 5G System (5GS)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 23.501, 2024, version 18.5.0.

[18] 3GPP, "User Equipment (UE) radio access capabilities," Technical Specification (TS) TS 38.306, Dec. 2025.

[19] D. Campoy, I. Vilà, J. Pérez-Romero, and O. Sallent, "Digital twin as a service for 6g radio access networks: Functional model and key challenges," in *2025 31st International Conference on Telecommunications (ICT)*, 2025, pp. 1–7.

[20] 3GPP, "NR; Physical layer procedures for data," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214, 2024, version 18.2.0.