

Article

Improving 5G User Plane Function Performance via Access Control Rule Distribution

Anne-Gaëlle Calandre ^{1,*}, David Espes ² and Johanne Vincent ³¹ Ernst and Young Advisory, EY, La Défense FR, Tour First, 1 Pl. des Saisons, 92400 Courbevoie, France² Laboratoire des Sciences et Techniques de L'Information de la Communication et de la Connaissance, Université de Bretagne Occidentale, Technopole Brest-Iroise, Cedex 3, 29238 Brest, France; david.espes@univ-brest.fr³ Laboratoire des Sciences et Techniques de L'Information de la Communication et de la Connaissance, Institut Mines-Télécom Atlantique, du Technopôle, 29280 Plouzané, France; johanne.vincent@imt-atlantique.fr

* Correspondence: ag.calandre@gmail.com or anne.gaelle.calandre@fr.ey.com; Tel.: +33-643-702-873

Abstract

The deployment of 5G technology represents a significant advancement in telecommunications, offering unprecedented speed, connectivity, and innovation opportunities. However, this progress comes at a significant cost for Public Land Mobile Network (PLMN) operators, who face challenges in meeting high Quality of Service (QoS) standards for optimal user experience while ensuring appropriate levels of security. This paper addresses the joint optimization of latency and resource consumption under security constraints within 5G networks, focusing on the Packet Data Unit (PDU) session path to ensure compliance with security and latency requirements. We propose an innovative approach in which access control rules are distributed across User Plane Functions (UPFs) in the network. The optimization problem has been formulated as a mixed integer linear programming (MILP) problem that aims to minimize round-trip latency and operational costs for PLMN operators. We evaluate the performance of our model using a discrete event network simulator (NS3). The simulation results demonstrate the effectiveness of our approach, particularly in scenarios with stringent latency requirements. Latency is reduced, and a lower session drop rate is maintained, especially in conditions of network congestion. These findings emphasize the importance of considering both QoS and security in the design of next-generation 5G networks.

Keywords: 5G; user plane function; access control; mixed integer linear programming; public land mobile network; network simulator 3



Academic Editors: Kostas Lampropoulos, Alexios Lekidis and Paris Kitsos

Received: 30 April 2026

Revised: 13 June 2026

Accepted: 25 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The focus of next-generation mobile communications, including future versions of the currently deployed 5G networks, is to provide seamless communication for a diverse set of traffic. With the promise of delivering unprecedented speed, ultra-low latency, and broad connectivity without compromising security, the deployment of 5G networks introduces a variety of challenges.

5G networks are designed to support three primary service categories defined by the International Telecommunication Union Radiocommunication Sector (ITU-R): enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC). The eMBB service class aims to provide high data rates and extensive coverage for bandwidth-intensive applications, such as

high-definition video streaming, virtual reality, and augmented reality. In contrast, mMTC is intended to enable large-scale connectivity for low-power Internet of Things (IoT) devices. mMTC is characterized by relaxed latency requirements and sporadic traffic patterns. URLLC represents the most demanding service category, requiring end-to-end latencies on the order of 1 ms while simultaneously ensuring extremely high reliability. Achieving Quality of Service (QoS) guarantees across these heterogeneous service categories within a single 5G New Radio (NR) system is a well-recognized challenge [1]. At the Radio Access Network level, factors such as channel variability, interference, and the coexistence of services with conflicting resource requirements make it particularly difficult to provide simultaneous latency and reliability guarantees. Beyond the Radio Access Network (RAN), the 5G core network introduces further challenges, as security mechanisms such as authentication, encryption, and access control impose additional processing overhead that directly impacts end-to-end latency, particularly for URLLC sessions.

In communication networks, and by extension in 5G networks, security and Quality of Service have traditionally been studied as separate issues. QoS refers to a qualitative assessment of various characteristics of the network, such as bandwidth, throughput, end-to-end delay, jitter, bit error rate, and packet loss. On the one hand, QoS is a central concern in the deployment of 5G networks. The technical specifications set by the 3GPP (3rd Generation Partnership Project) impose strict requirements regarding QoS, and as a result, extensive research has been conducted on this subject, particularly to address specific use cases such as vehicular ad hoc networks (VANETs) [2] and 5G healthcare applications [3]. On the other hand, security within 5G networks is also a critical issue, and numerous studies have proposed different solutions to protect communications [4,5]. Several approaches exist, including those directly proposed by the 3GPP, which outlines six security domains: network access security, network domain security, user domain security, application domain security, Service-Based Architecture (SBA) domain security and visibility, and configurability of security [6]. Additional research aiming to reinforce the foundational requirements established by the 3GPP, including strengthening authentication protocols [7] and implementing an innovative Network Anomaly Detection (NAD) architecture [8] has been conducted.

Each type of service has specific requirements for both security and QoS to achieve optimal performance. However, meeting both security and QoS demands simultaneously can be challenging due to limited network resources and substantial costs for the Public Land Mobile Network (PLMN) operators. Consequently, it is essential to consider the interaction between security and QoS. To achieve a stable QoS, security has to be carefully considered to resist diverse attacks and prevent traffic congestion. Inadequate network security can make the system susceptible to attacks that impact QoS, such as the distributed denial of service (DDoS) attacks or jamming attacks. However, improving security might increase latency and delay and thus affect QoS, since authentication, encryption and access control all have a direct impact on latency [9]. This paper focuses specifically on latency as the QoS factor, since it is the primary aspect affected by security measures.

In the 5G core network, the user plane is responsible for routing data traffic between the access network and the data network and thus plays a central role in delay. Traditionally, the user plane has been deployed centrally, utilizing long-distance backhaul links, which results in high end-to-end latency. This centralized approach falls short of meeting the ultra-low latency demands of emerging 5G applications. One way to address this challenge is to distribute the deployment of User Plane Functions (UPFs) at the network edge, bringing them closer to the users. This strategy significantly reduces end-to-end latency [10].

Most existing approaches optimize security and QoS (particularly latency) separately, which is inadequate for emerging ultra-low latency services. Centralized user plane deployments and security-agnostic optimizations cannot meet both stringent latency and

robust security requirements under the strict cost and resource constraints faced by PLMN operators. This gap motivates a joint, security-aware strategy with distributed UPFs and integrated access control to simultaneously minimize latency and operator cost while maintaining the required security levels. In this paper, we investigate the joint optimization of latency and resource consumption. Our goal is to optimize the round-trip latency and reduce the resource consumption for the PLMN operator while ensuring that the Packet Data Unit (PDU) session respects the latency requirement and adheres to a minimum set of access control (AC) rules. The problem is formulated as a mixed integer linear programming (MILP) problem. The main contributions of this paper are summarized as follows:

1. We formulate the problem as a MILP problem; the primary goal of our model is to optimize both the latency and PLMN operator cost while ensuring that the path followed by the PDU session respects both the latency requirement and the security constraints;
2. We implemented our model in a discrete event network simulator (NS3) to observe its performance;
3. We validate the effectiveness of our model by comparing it to a second approach in which the AC rules are not distributed along the PDU session path.

The remainder of this article is organized as follows: In Section 2, we present an overview of relevant research regarding the trade-off between QoS and security on the one hand and the UPF selection problem on the other. Section 3 introduces the system model and the notation used. In Section 4, we present our solution to solve the system model. Section 5 presents the parameters of the simulation in NS3, and Section 6 presents the results of the conducted simulations. Finally, Section 7 presents our conclusions.

2. Related Work

In the following paragraphs, we will briefly review the literature related to the trade-off between QoS and security and UPF selection and function.

2.1. QoS and Security Trade-Offs

Initially treated separately, the joint consideration of Quality of Service (QoS) and security metrics in 5G networks has emerged as an important research challenge.

The authors in [11] address the Intrusion Detection Mechanism (IDM) selection problem: monitoring the system at higher detection levels requires sophisticated IDMs, which consume more resources, including network bandwidth and computational resources, resulting in increased monitoring overheads. Their model considers various factors, including network traffic patterns, bandwidth, potential threat vectors, and the computational overhead of different IDM configurations. In ref. [12], the authors developed a joint security and QoS utility function based on the Cobb–Douglas production utility model. They demonstrated that increasing the detection level of the IDM enhances the security while simultaneously negatively impacting the QoS, since more resources (computational, bandwidth and monetary) are utilized for IDM processing. These articles investigated a different solution from the one we present in this paper. In our approach, security is enhanced through the application of various access control rules, while optimization is achieved by distributing these rules across the UPFs. In ref. [13], the authors proposed a model in which QoS and security are modeled as a stochastic game. They use a flexible game-theory approach to determine multilevel security based on the trade-off between the encryption algorithms, the key length as the security parameters, and the network throughput as a QoS parameter. The approach examined by the authors in their work is partially similar to ours, as it is based on optimizing resource usage across various eNBs. However, it differs in other aspects: access control rules are not used as a security parameter, and only

the capabilities of the evolved NodeB (eNBs) are considered. In this paper, we consider latency as a QoS metric and access control as the security parameter. Moreover, we consider distributing access control rules across various UPFs, an approach not employed in any of the existing studies.

The authors in [14] formalized a mathematical model for latency-optimal on-path allocation of virtual network security functions within an edge network infrastructure. Their objective is to provision and allocate security services by composing virtual function chains according to specific QoS needs of users and security policies. The security policies included: the type (e.g., firewall, Intrusion Detection System, ACL, etc.) that should be deployed for a specific class of applications and their order. They integrate a flow matrix in a context-aware placement problem of a security function to minimize path latency. The problem is formalized as an integer linear programming (ILP) and has been solved using a commercial integer linear programming solver. This approach, while not directly applied to 5G networks, bears similarities to our own; however, several distinctions should be noted: their work focuses on security function placement, and during their simulation, they assumed a chain of two security functions per user, and the PLMN cost was not considered. In contrast, our approach focused on the optimal path for the PDU session and also aims to reduce costs for PLMN operators.

While there is limited research on 5G networks, the trade-off between security and QoS has been extensively studied in virtualized environments [14–16]. The approaches and issues addressed in these works vary. In ref. [15], the authors studied the optimization of allocation and life cycle management of virtual network functions in various environments. They proposed an orchestration system that considers the QoS of the system (including end-to-end delay, bandwidth, jitter, etc.) and the security level (via security network functions such as firewalls, proxies, etc). They developed a custom heuristic algorithm to find the near-optimal configuration, including the use of machine learning-based algorithms and a Hidden-Markov-Model-based optimization algorithm. In reference [16], virtual network security function chains handle network services that may require specific conditions, like maximum bandwidth or minimum delay. The authors introduce a security service embedding approach, designed to optimize resource use and deploy the virtual security function chain effectively, based on individual application security needs and the operator's strategy.

In ref. [17], the authors studied the impact of access control rule processing time on Quality of Service. They explore the trade-off between performance and security in high-speed network routers to determine the extent to which router performance is impacted when security mechanisms are activated and whether such performance degradation is acceptable for real-time applications in high-speed networks. To address these questions, queuing models of a router node, equipped with single-core and quad-core processors, are analyzed. The study focuses on the impact of the Access Control List (ACL) security mechanism, which causes the router to become a network bottleneck by degrading the router's forwarding speed. Router performance is evaluated in terms of the mean response time, server utilization, mean queue length, and packet loss probability. Their results show that the ACL security application has an adverse effect on the performance of the router. The activation of ACL security negatively impacts router performance by increasing the response time and packet loss rate.

2.2. User Plane Selection and Function

Recently, several studies have focused on the deployment of the 5G core network user plane, such as [18], where the authors study the dynamic UPF placement and session mapping reconfiguration problem. They propose an integer linear program (ILP) and a

heuristic algorithm to determine the best UPF placement and PDU session mapping setup to ensure compliance with 5G Quality of Service requirements. In ref. [19], the authors examine the joint optimization of traffic routing and UPF placement. Their goal is to reduce the overall cost of UPF deployment, user plane latency, and energy usage. The problem is formulated as a mixed integer linear programming (MILP) problem and is NP-hard. They suggest a Benders decomposition-based technique to lower the computing cost of solving this issue. In ref. [20], the authors investigate the minimization of latency within cost constraints by jointly deploying edge servers and UPFs in a 6G scenario. However, none of these works takes into account the security requirements and the latency induced by those security mechanisms.

None of the previously cited works simultaneously considers QoS, PLMN costs, and security. Furthermore, the existing studies do not address the role of UPFs in security, particularly their ability to filter PDU sessions using ACL. Unlike the existing literature, our work investigates the optimal PDU session path under security and latency constraints while ensuring minimal costs for PLMN. In this context, we propose an innovative approach in which UPFs are utilized to enforce AC rules throughout the entire PDU session path.

2.3. Security Service Function Chaining

Security Service Function Chaining (S-SFC) and security-aware routing have recently attracted significant attention as mechanisms for enforcing security policies along data paths in virtualized networks [21–26].

In S-SFC-based approaches, security is enforced through an ordered chain of dedicated virtual security functions instantiated as independent virtual nodes and inserted into the data path. Each security function constitutes a separate processing entity that must be traversed sequentially. The authors of [23] formulate this problem as a MILP problem for the latency-sensitive placement of virtual security functions and show that, even under optimized placement, the introduction of dedicated security nodes inherently increases end-to-end latency. In ref. [22] the authors further investigate the reliability of S-SFC deployments, highlighting that each additional security function represents an independent failure-prone component with its own resource requirements. More broadly, these approaches model security enforcement through a dedicated service chain that must be constructed and embedded separately from the routing problem.

In contrast, our approach deploys AC rules within UPFs that already exist in the network topology. Consequently, security enforcement becomes an intrinsic property of the routing decision, rather than a separate service-chain design problem. The security requirement is modeled as a coverage constraint over the routing variables, while the set of AC rules applied to a session is determined directly by the UPFs selected along its path. Security enforcement and routing are therefore jointly optimized within a single formulation. To our knowledge, this level of integration is not considered in the S-SFC and security-aware routing literature reviewed in this work.

Finally, the reviewed S-SFC and security-aware routing literature has been developed primarily for generic Network Function Virtualization (NFV) infrastructures and does not account for the specific architectural characteristics of 5G networks, where UPFs already provide native packet inspection and filtering capabilities as defined by the 3GPP. Consequently, rather than leveraging the native packet inspection and filtering capabilities already present in UPFs, these approaches introduce dedicated virtual security functions as separate network nodes, which incur additional deployment and operational costs. Furthermore, none of the reviewed works considers the PLMN operator cost as an optimization objective.

3. System Model and Mathematical Representation

This section provides a concise overview of 5G UPFs, followed by the system model. The mathematical model of our system is presented herein, aiming to generate PDU session paths, ensuring compliance with security and latency requirements, and minimizing operator costs. The notations utilized in this article are detailed in Table 1.

Table 1. Notations.

Notation	Description
Sets	
N	Set of nodes
N_e	Set of user nodes
N_a	Set of access nodes
N_{UPF}	Set of nodes hosting a UPF
U	Set of UPF
E	Set of physical link
P	Set of paths between physical node
A	Set of access control rules
S	Set of PDU session
K	Set of network slice
A_s^{req}	Predefined set of requested AC rules for PDU sessions $s \in S$
A^{PLMN}	Predefined set of general network-wide AC rules applicable to all sessions regardless of their slice
$A_{u,n}$	Set of AC rules deployed on UPF $u \in U$ hosted in node $n \in N$
Parameters	
d_{DN}	Processing time of DN
l_{RAN}	Latency of RAN trip
R_n	Capacity (e.g., CPU, memory, etc.) of node $n \in N$
$\beta_{u,v}$	Bandwidth capacity of physical link $(u, v) \in E$
β_s	Requested bandwidth for PDU session $s \in S$
$L_{u,v}$	Latency of physical link $(u, v) \in E$
L_s^{req}	Latency required for PDU session $s \in S$
$C_{u,n}$	Processing capacity of UPF $u \in U$ hosted in node $n \in N$
O_u	Operating cost of UPF $u \in U$
C_s	Requested capacity for PDU session $s \in S$
T_a	Processing time for rule $a \in A$
R_u	Requested resources for UPF $u \in U$
$T_{u,n}$	Processing time of UPF $u \in U$ hosted in node $n \in N$
Binaries	
$y_{u,v}^s$	1 if PDU session $s \in S$ transits via physical link $(u, v) \in E$
$x_{u,n}$	1 if UPF $u \in U$ is deployed on node $n \in N$
$z_{s,u}^n$	1 if UPF $u \in U$ hosted in node $n \in N$ processes PDU session $s \in S$
$w_{a,u}^n$	1 if rule $a \in A$ is present in UPF $u \in U$ hosted in node $n \in N$
o_u^n	1 if UPF $u \in U$ hosted in node $n \in N$ processes at least one PDU session

3.1. User Plane Function in 5G Networks

A 5G network, see Figure 1, consists of two distinct parts: the wireless Radio Access Network, which includes User Equipment (UE) and the g-NodeB (gNB), and the wired packet core network composed of various network virtual functions. In the 5G network core's control plane, the Access and Mobility Function (AMF) handles registration and mobility management, while the Session Management Function (SMF) oversees the UE's data sessions. SMFs are responsible for selecting, instructing, and managing the UPFs associated with PDU sessions during their lifetimes. The data plane is formed by the

UPFs, which combine functionalities of traditional serving and packet gateways. The UPFs' primary function is to process data plane packets between the access network and the data networks, but they also perform other functionalities, such as packet routing and forwarding, traffic steering, QoS handling, packet inspection, and lawful interception. Moreover, PDU sessions can have different characteristics and demands according to their topology and service demands.

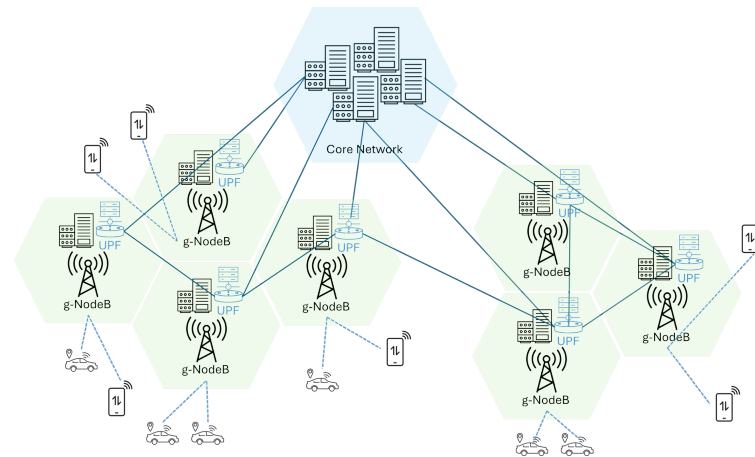


Figure 1. Simplified 5G network architecture, illustrating the distinction between the RAN and the 5G core network.

Due to their position and flexibility, UPFs play a central role in QoS, particularly in terms of latency. They also serve as the entry point into the wired packet core network, giving them an important role in security, especially through traffic filtering functions. In this paper, we explore the possibility of distributing access control rules across all UPFs along the PDU session path to optimize latency without compromising security.

As defined within the 3GPP standards, in 5G architectures, distributed UPFs can operate as Uplink Classifiers (ULCLs) or Branching Points, intercepting uplink traffic at the edge, evaluating packet headers against localized rules, and routing approved data directly to local data networks or edge computing platforms, thereby bypassing the core network entirely [27,28]. This architecture is central to our approach: by distributing AC rules across UPFs positioned along the PDU session path, enforcement decisions are made as close to the traffic source as possible. Furthermore, UPFs at the edge natively inspect the QoS Flow Identifier (QFI) carried within GTP-U (GPRS Tunneling Protocol-User) packet headers, which allows them to immediately identify high-priority, latency-sensitive flows such as URLLC sessions and enforce the corresponding resource allocation policies without waiting for centralized core processing [27]. From a security perspective, the proximity of UPFs to the network edge means that access control decisions such as gating, rate-limiting, and dropping of unauthorized flows occur with minimal propagation delay [27,29]. This prevents rogue or unauthorized traffic from consuming backhaul bandwidth before being identified. Our work exploits precisely these native capabilities: rather than introducing dedicated security nodes as in S-SFC approaches, we leverage the existing packet inspection and filtering functions of UPFs as defined by the 3GPP, integrating access control enforcement directly into the routing decision.

3.2. Scope of the Security Model

Although 5G networks encompass a wide range of security mechanisms, not all contribute to packet forwarding latency along the PDU session path. Authentication, as specified in 3GPP TS 33.501 [30], is triggered exclusively at initial UE attachment and upon inter-AMF handover events, not during an established PDU session; therefore, its

overhead is a one-time control-plane cost. The encryption overhead is small and fixed: IPsec correctly configured on the N3 interface introduces a user plane per-packet overhead of approximately 55 μ s [31], which is path-independent and cancels out in any relative comparison between routing strategies. Key management follows a similar argument, as session keys are derived once at PDU session establishment and renewed only at low frequency under the operator policy [30]. Stateful firewall processing is captured within the AC rule processing time. Regarding Deep Packet Inspection (DPI), while inline DPI introduces measurable processing overhead when rules are actively applied per packet, the complex AC rules already modeled in our framework capture precisely this inline inspection cost.

3.3. Network Slicing and Slice-Aware Access Control

Network slicing is a key feature of 5G networks that enables the simultaneous operation of multiple logically isolated virtual networks over a shared physical infrastructure [27]. Each network slice is identified by a Single Network Slice Selection Assistance Information (S-NSSAI), composed of a Slice/Service Type (SST) and an optional Slice Differentiator (SD) [27]. The S-NSSAI associated with a PDU session is made available to the UPF via the N4 interface [29], enabling slice-aware traffic handling and policy enforcement directly within the user plane. A single UE may simultaneously belong to up to eight network slices within a given PLMN, each with distinct QoS and security requirements.

From a security perspective, each network slice may be associated with a distinct set of access control policies. In this context, we extend our model to the network slicing paradigm: the operator defines a set of mandatory AC rules for each slice, and the total set of rules that a PDU session must satisfy is the union of the general network-wide rules and the rules specific to the slice to which the session belongs.

Let A_s^{req} denote the set of mandatory access control (AC) rules associated with a PDU session $s \in S$, where S is the set of all PDU sessions. Let K denote the set of network slices supported by the PLMN. The mandatory AC rules for a session are composed of a set of network-wide rules and a set of slice-specific rules:

$$A_s^{req} = A^{PLMN} \cup A_s^k \quad \forall s \in S, \quad (1)$$

$$A^{PLMN} \subset A, \quad (2)$$

where A denotes the set of all access control rules, A^{PLMN} is the predefined set of AC rules enforced across the entire PLMN and is therefore applicable to every PDU session, and A_s^k is the set of AC rules specific to the network slice $k \in K$ to which session $s \in S$ belongs.

3.4. System Model, Abbreviations and Acronyms

In this paper, the 5G network is modeled as an undirected graph $G(N, E)$, where N is the set of nodes and E is the set of edges. The set of nodes is composed of three subsets: user nodes N_e , UPF access nodes N_a , and intermediary UPF nodes N_{UPF} . The set of deployed UPFs is denoted U , $U = N_{UPF} \cup N_a$. The nodes and edges have pre-set maximum capacity values that are used as boundaries for the traffic. R_n represents the resources (e.g., CPU and memory) of node $n \in N$; $\beta_{u,v}$ represents the bandwidth capacity of the physical link $(u, v) \in E$; $L_{u,v}$ represents the latency of the physical link $(u, v) \in E$. A node can host multiple UPFs. Each UPF, independently of its type, has a processing capacity $C_{u,n}$ and a processing time $T_{u,n}$. The processing time is calculated based on the number of access control rules executed by the UPF and their complexity. The set of AC rules is denoted by A . Each rule has a complexity level that impacts the processing time of the rule. The set of rules deployed on the UPF $u \in U$ hosted on node $n \in N$ is indicated by $A_{u,n}$.

We consider the following scenario: a set of devices (e.g., UEs, smart vehicles, etc.) is connected to the access network through the nearest access node and requests one or multiple PDU sessions. Each PDU session, $s \in S$, requires a processing capacity C_s , a bandwidth β_s , an end-to-end delay requirement L_s , and a security requirement defined as a set of predetermined mandatory AC rules A_s^{req} .

The following equations refer to the cost of resources to route a PDU session on the network. We assume that a node can host multiple UPF instances and that there is a maximum of one physical link between two nodes. Equation (3) represents the bandwidth utilization for a given link, while Equation (4) addresses the resource consumption of a specific node, and Equation (5) describes the resource consumption of a UPF.

$$\beta_{u,v}^{used} = \sum_{s \in S} \beta_s \times y_{u,v}^s \quad \forall (u,v) \in E \quad (3)$$

$$R_n^{used} = \sum_{u \in U} R_u \times x_{u,n} + \sum_{u \in U} O_u \times o_u^n \quad \forall n \in N \quad (4)$$

$$C_{u,n}^{used} = \sum_{s \in S} C_s \times z_{s,u}^n \quad \forall u \in U; \forall n \in N \quad (5)$$

The subsequent equations address the various delays introduced by the routing and processing times for a PDU session. Equation (6) represents the latency caused by transport across different links. Equation (7) specifies the delay introduced by a single UPF. Equation (8) addresses the delay resulting from all UPFs processing a PDU session.

$$C_{rout}^s = \sum_{(u,v) \in E} L_{u,v} \times y_{u,v}^s \quad \forall s \in S \quad (6)$$

$$T_{u,n} = \sum_{a \in A} T_a \times w_{a,u}^n \quad \forall u \in U; \forall n \in N \quad (7)$$

$$C_{proc}^s = \sum_{u \in U} \sum_{n \in N} T_{u,n} \times z_{s,u}^n \quad \forall s \in S \quad (8)$$

The QoS is measured in terms of service latency, which is expressed as the combination of UPF and data network (DN) processing time and propagation delays between consecutive nodes.

To estimate the latency, we relied on the results presented in [32].

In this regard, the following equation expresses the end-to-end round-trip latency. Here, we assumed that the processing time in the DN d_{DN} is fixed and that the session would follow the same path in both directions: from the UE to the DN and from the DN back to the UE.

$$L_s = 2 \times (C_{rout}^s + C_{proc}^s + l_{RAN}) + d_{DN} \quad \forall s \in S \quad (9)$$

Each PDU session adheres to all access control rules enforced by the UPFs that process it. Let A_s denote the set of AC rules effectively applied to PDU session $s \in S$ along its path. A rule $a \in A$ is applied to session s if, and only if, there exists at least one UPF $u \in U$, hosted on node $n \in N$, such that rule a is deployed on that UPF ($a \in A_{u,n}$) and that UPF processes session s ($z_{s,u}^n = 1$). Formally:

$$a \in A_s \iff \exists u \in U, n \in N : a \in A_{u,n} \wedge z_{s,u}^n = 1 \quad \forall s \in S. \quad (10)$$

Equivalently, using set notation:

$$\mathcal{A}_s = \bigcup_{\substack{u \in \mathcal{U}, n \in \mathcal{N} \\ z_{s,u}^n = 1}} \mathcal{A}_{u,n} \quad \forall s \in \mathcal{S}. \quad (11)$$

3.5. Multi-Objective Optimization Problem

Given an initial network setup, the objective is to determine the optimal path for each PDU session $s \in \mathcal{S}$ that simultaneously minimizes round-trip latency L_s and node resource consumption R , while respecting QoS requirements, security constraints, and resource limitations. The two objectives, latency minimization and resource consumption minimization, are, in general, conflicting: distributing load across more UPFs reduces per-node utilization, but may increase propagation latency, while concentrating traffic on fewer, centrally located UPFs reduces routing latency but increases resource consumption and may create bottlenecks.

We adopt a weighted-sum scalarization approach, which converts the two objectives into a single scalar objective function:

$$\min \quad \alpha \times \sum_{s \in \mathcal{S}} \sum_{n \in \mathcal{N}} L_s + (1 - \alpha) \times R_n^{used}, \quad (12)$$

where $\alpha \in [0, 1]$ is a weighting coefficient that allows one objective to be favored over the other. In the simulations presented in Section 5, we set $\alpha = 0.5$ to give equal weight to both objectives. This value reflects the assumption that latency and operational cost are of equal importance to the PLMN operator.

This is subject to:

$$A_s^{req} \subseteq A_s \quad \forall s \in \mathcal{S}, \quad (13)$$

$$L_s \leq L_s^{req} \quad \forall s \in \mathcal{S}, \quad (14)$$

$$\beta_{u,v}^{used} \leq \beta_{u,v} \quad \forall (u, v) \in E, \quad (15)$$

$$R_n^{used} \leq R_n \quad \forall n \in \mathcal{N}, \quad (16)$$

$$C_{u,n}^{used} \leq C_{u,n} \quad \forall n \in \mathcal{N}; \forall u \in \mathcal{U}. \quad (17)$$

Equation(18) requires that the set of mandatory rules A_s^{req} is a subset of A_s , i.e., every required rule must be covered by at least one UPF on the session path:

$$\forall a \in A_s^{req}, \exists u \in \mathcal{U}, n \in \mathcal{N} : a \in A_{u,n} \wedge z_{s,u}^n = 1 \quad \forall s \in \mathcal{S}. \quad (18)$$

Equation (19) can be rewritten as:

$$\sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} w_{a,u}^n \times z_{s,u}^n \geq 1 \quad \forall s \in \mathcal{S}, \quad \forall a \in A_s^{req}, \quad (19)$$

where $w_{a,u}^n$ is the binary variable defined in Table 1, which is equal to one if rule a is present on UPF u hosted on node n . This constraint ensures that at least one UPF processing session s enforces each mandatory rule, while remaining fully compatible with the MILP branch-and-bound resolution described in Section 4.

Equations (15)–(17) are related to the resource limitations of the network. Specifically, Equation (16) ensures that the UPF instances deployed at the nodes do not exceed the available resources of those nodes (e.g., CPU and memory). Similarly, Equations (15) and (17) enforce that the selected UPFs and links utilized to route a session possess sufficient processing capacity and bandwidth to meet their demands.

4. Proposed Solution

The considered optimization problem jointly selects a path for each PDU session and allocates node and link resources so as to minimize end-to-end latency and resource consumption under QoS, security and capacity constraints. This type of decision problem lies in the same class as classical NP-complete combinatorial optimization problems cataloged in [33]. Formally, even if the objective is simplified to deciding whether there exists a feasible path and UPF placement satisfying Equations (13)–(17) and given upper bounds on latency and resource usage, the problem remains computationally intractable. Minimizing the number of possible nodes while keeping all flows feasible under chaining and capacity constraints is NP-hard [34], and joint placement-allocation of VNF nodes with budget and capacity constraints is also NP-hard [35]. Since the proposed model subsumes these problems as special cases, it is at least as hard. Therefore, following the standard complexity theory framework of Garey and Johnson [33], we can deduce that our problem is also NP-hard, and no polynomial-time algorithm exists to solve it unless $P = NP$.

We opted for a MILP-based algorithm using a branching approach. This choice ensures that the entire set of constraints, particularly those related to security, is adhered to. Thus, we have the assurance that the security level of the PDU sessions, as well as the Quality of Service criteria, are systematically respected.

Linear programming (LP)-based branch-and-bound techniques for MILP involve solving LP subproblems derived by relaxing certain constraints, most specifically, integrality constraints. This relaxation provides bounds on the objective function and demonstrates infeasibility in a set of constraints when applicable. A node is discarded if the objective value from the LP relaxation is worse than the best integer solution found so far or if the LP subproblem is infeasible. Branching occurs when the LP relaxation solution fails to satisfy all the constraints of the original problem. If the relaxed solution satisfies all constraints and improves upon the best feasible solution identified so far, the best feasible solution is updated. The search ends when it is proven that there can be no better solution than the current best feasible solution. The efficiency of MILP methods is influenced by the size of the LP subproblems and, more critically, by the gap between the objective value of the best feasible solution (optimum) and the value obtained from the initial LP relaxation [36,37].

5. Simulations

This section presents a series of experiments conducted to evaluate the performance of the proposed solution. We begin by introducing key aspects of the simulation environment. Next, we assess the effectiveness of the solutions. Finally, we analyze the performance of the solution.

5.1. Simulation Configuration

For the network topology we divided a region-wide network into 14 hexagons. Each hexagon contains:

- 50,000 UE (user nodes);
- 100 gNB (access nodes);
- 200 UPF (intermediary UPF).

Each hexagon has at least one link to an intermediary UPF node in each neighboring hexagon. Within each hexagonal region, the positions of UEs, gNBs, and UPF nodes are drawn independently from a uniform random distribution over the hexagonal area (using the `numpy.random.uniform` function). This spatial modeling choice is consistent with the reference simulation methodology used by the 3GPP in TR 36.873 [38].

UEs are connected to the gNB via the RAN. By default, the nearest gNB is selected; however, if this gNB has reached its maximum capacity and cannot accommodate additional users, the next closest gNB is chosen. The process continues in this manner, as necessary. Each gNB node also contains a UPF.

As explained in the 3GPP, the maximum number of PDU sessions that a UE can establish in a PLMN is limited by whichever is the lowest: the maximum number of PDU session IDs allowed by the protocol (as specified in 3GPP TS24.007) or the PLMN's maximum number of PDU sessions allowed. In this simulation, we assume that the PLMN restricts each user to a maximum of eight PDU sessions. A UE can be connected to only one gNB at a time. We assumed that the duration of a PDU session ranges from 5 to 60 s.

Based on work conducted in [39], a propagation delay of 5 μ s per kilometer is assumed for fiber optic links. For the RAN (Radio Access Network) trip (between UE and gNB), a fixed delay of 400 μ s is applied.

The parameters used for the simulation are presented in Table 2 below.

Table 2. Simulation parameters.

Notation	Description	Value
N_e	Set of user nodes	700,000
N_a	Set of access nodes	1400
N_{UPF}	Set of nodes hosting a UPF	4200
R_{AC}	Capacity of an access node (CPU)	2400
R_{IN}	Capacity of an intermediary node (CPU)	1800
$C_{(UPF)}$	Processing cap. of a UPF	[600, 800]
$\beta_{(u,v)}$	Bandwidth cap. of optical fiber (Gbps)	100
S_{UE}	Number of PDU sessions per UE	[0–8]
d_{dn}	Processing time of DN (μ s)	500
$L_{(u,v)}$	Prop. delay of links (μ s/km)	5
l_{RAN}	Prop. delay of the RAN (μ s)	400
β_s	Bandwidth requirement (Mbps)	[10, 25, 50]
C_s	CPU processing demands of a PDU session	[0.1, 0.3, 0.5]
O_u	Operating cost of a UPF (CPU)	10

5.2. 5G Traffic

We consider three 5G use cases:

- **Web Browsing:** This remains one of the most prevalent use cases for 5G networks and does not typically require low latency. As noted in [40], human physiological reaction times mean that further reductions in latency, leading to faster web response times, do not noticeably enhance the user experience, as the difference is imperceptible. Even early technologies such as 802.11b Wi-Fi and 3G cellular systems were generally able to meet these latency requirements, ensuring a seamless real-time browsing experience for users. A latency of 140 ms is considered acceptable for this use case.
- **Intelligent Transportation Systems (ITS):** These rely on ultra-reliable low-latency communication (URLLC) to support autonomous driving and optimize road traffic management. ITS encompasses a variety of use cases, including autonomous driving, road safety, and traffic efficiency services, each with specific requirements. Autonomous vehicles must coordinate their actions, such as platooning and overtaking, to ensure safety and efficiency. For example, the pre-sense crash warning has a minimum requirement of 20–50 ms of round-trip latency [41]. In our simulation, for this use case, we will adopt a maximal latency of 30 ms.

- **Healthcare Applications:** Low-latency communication enables advanced medical services, including remote physical examinations, robotic-assisted surgeries, and real-time patient monitoring. As stated in ref. [42], haptic feedback demands a lower latency compared to visual feedback, with latencies exceeding 20 ms having a minimal impact on performance.

During the simulations, traffic distribution is as follows: 70% corresponds to the first use case, while the second and third use cases each account for 15%.

5.3. Access Control Rules

We have categorized the AC rules into two groups, based on their level of complexity:

- **Basic AC Rule:** These are simple filtering rules based on elements such as the source IP, destination IP, port, protocol used, and other information present in the packet header. Drawing from the data presented in the comparative study [43], we can conclude that the processing time for a basic filtering rule (source IP filtering in this comparison) for a 256-byte packet is:

$$t = 69 \mu\text{s}.$$

- **Complex AC Rule:** These rules require DPI as well as header inspection. In ref. [44], the authors conducted an experimental evaluation of the latency induced by an advanced DPI filtering rule on the firewall:

$$T = 835 \mu\text{s}.$$

Access control rules, both simple and complex, are aggregated by action group. For instance, simple rules may involve any-any filtering or rules to filter specific protocols (such as BitTorrent). Complex rules, on the other hand, pertain to application-level filtering tailored to the use cases presented in the previous section.

The specific AC rules assigned to each use case reflect the security requirements typical of each service category. For the web browsing use case, basic rules include source and destination IP filtering, port-based filtering, and protocol filtering (e.g., blocking peer-to-peer protocols such as BitTorrent). Complex rules include application-level content filtering and URL categorization, enabled through DPI, which operates at the application layer to inspect packet payloads beyond header information [44]. For the Intelligent Transportation Systems use case, basic rules focus on whitelisting authorized vehicle identifiers and filtering based on the protocol type, allowing only V2X-specific protocols (Vehicle-to-Everything). Complex rules include DPI-based inspection to verify message integrity and detect anomalous commands that could indicate spoofing or impersonation attacks, which represent well-documented threats in V2X environments [45,46]. For the healthcare applications use case, basic rules enforce strict source IP whitelisting to ensure that only authorized medical devices can initiate sessions, consistent with the access control principles defined in 3GPP TS 33.501 for securing sensitive user plane data. Complex rules include DPI-based inspection of application payloads to detect unauthorized access to patient data streams and to verify compliance with medical device communication protocols [3,44].

In the proposed simulation, we consider a set of 20 basic and 30 complex AC rules. For each use case presented below, we specify a minimum number of basic and complex rules to which the PDU session must adhere. The selection of specific rules is determined randomly at the beginning of each session.

5.4. Solver Selection

To address our problem, we examined three Python 3.10.12.-compatible solvers: two commercial options, Gurobi and CPLEX, and one open-source option, CBC. These solvers are widely used for optimization problems but vary significantly in terms of performance, features, and licensing. Gurobi and CPLEX generally outperform CBC [47], particularly for large-scale and complex problems, offering faster and more efficient solutions. CBC is typically more suitable for small to medium-sized problems, whereas Gurobi and CPLEX excel with large-scale and complex tasks. Due to Gurobi's superior performance in handling large-scale optimization problems, compared to CPLEX [48], we selected it as the solver for our problem.

After generating the topologies, defining the model and its constraints, and obtaining the optimal path for each PDU session from the Gurobi solver, all this information is transferred to NS3 for simulation.

5.5. NS3

NS3 (Ns-3.42) is a C++-based discrete event network simulator that provides a Python scripting API. To support our simulation, we adhered to the NS3 standard architecture and developed custom helpers, applications, and a routing protocol based on existing modules.

For the routing protocol, we used an existing protocol, Nix-Vector Routing, which we adapted to our needs. Nix-Vector Routing is a static routing protocol that performs on-demand route computation. In the native version implemented in NS3, the route is calculated using a Breadth-First Search (BFS) algorithm. When a packet is generated at a node for transmission, the route is calculated, and the nix vector is built. In our version, the nix-vector is no longer computed using BFS but is, instead, directly embedded in the packet. This nix-vector represents the optimal route for the given PDU session, as determined by the Gurobi solver.

To simulate the UEs and generate the traffic, we reused an existing traffic generator, OnOffApplication, and customized it. This traffic generator follows an On/Off pattern: after application :: StartApplication is called, "On" and "Off" states alternate. The duration of each of these states is determined by the onTime and offTime random variables. During the "Off" state, no traffic is generated. During the "On" state, Constant Bit Rate (CBR) traffic is generated. This CBR traffic is characterized by the specified "data rate" and "packet size". Our version of OnOffApplication has an m_nixVector attribute whose value is set to the optimized route returned by the Gurobi solver. In order to respect the NS3 logic, we also had to develop a custom helper OnOffHelperNixVector.

The purpose of these developments is to evaluate our proposal in a simulated environment that closely replicates the characteristics of a 5G network, allowing us to assess its relevance.

The pseudo-code of the simulation implementation is shown in Algorithms 1–3; the simulation comprises three main stages: the network topology creation (Algorithm 1), the optimal path computation (Algorithm 2) and the traffic simulation (Algorithm 3).

M_{Adj} is the adjacency matrix representing the graph $G(N, E)$. F_{vector} contains, for each user per PDU session, the optimal path pre-formed for the routing protocol, the packet source and destination.

The source code of the NS3 implementation, including the custom routing protocol, the modified OnOffApplication, and the OnOffHelperNixVector helper, is publicly available at: <https://github.com/AclD31/NS3-DARC> (access date: 15 June 2026).

5.6. Computational Performance and Scalability

The NP-hardness of the problem, established in Section 4, implies that exact optimal resolution becomes computationally intractable as the problem size grows. To address this, the MILP problem is not solved globally over the entire network. Instead, the optimization problem is first solved within a single hexagonal region; subsequently, the shortest path is used to reach the DN. Regarding convergence, Gurobi was configured with a relative MIP gap tolerance of 1×10^{-4} and a time limit of 2 s per subproblem. In all scenarios, the solver terminated within the time limit. The optimality gap remained below 1×10^{-4} across all runs, confirming that the solutions produced are near-optimal and that the weighted-sum objective defined in Equation (12) is minimized to within acceptable bounds.

The scalability of the framework is inherently bounded by the NP-hard nature of the underlying problem. For significantly larger topologies or dynamic re-optimization scenarios, heuristic or machine learning-based approaches could be considered as future extensions, as discussed in Section 7.

Algorithm 1 Network topology generation

Input: $H = 14$ hexagons;
Output: Graph $G(N, E)$, adjacency matrix M_{Adj} , rule sets $\{A_{u,n}\}$
for each hexagon h **do**
 Place 50,000 UEs, 100 gNBs, 200 UPFs using `numpy.random.uniform`
 Set node capacities based on node type
 for each UE $e \in N_e$ **do**
 Assign e to nearest gNB g^* with available capacity
 end for
 for each pair (u, v) within h with $\|pos(u) - pos(v)\| \leq d_{max}$ **do**
 Add edge to E ; set $L_{u,v} \leftarrow 5 \mu s \times d$, $\beta_{u,v} \leftarrow 100$ Gbps
 end for
end for
for each pair of neighboring hexagons (h, h') **do**
 Add at least one cross-border edge between border UPF nodes
end for
for each UPF $u \in U$ on node n **do**
 Sample $A_{u,n}$ from A_{basic} ($|=20$) and $A_{complex}$ ($|=30$)
end for
Build M_{Adj} from $G(N, E)$
Return $G, M_{Adj}, \{A_{u,n}\}$

Algorithm 2 MILP-based optimal path computation

Input: $G(N, E), \{A_{u,n}\}, \{R_n\}, \{L_{u,v}\}, \{C_{u,n}\}; \alpha = 0.5, \text{MIPGap } \varepsilon = 10^{-4}, \text{TimeLimit} = 2 \text{ s}$
Output: Flow vector F_{vector} (nix-vector path, source, destination per session)
for each UE $e \in N_e$ **do**
 Draw $n_s \sim \mathcal{U}(\{0, \dots, 8\})$ sessions; assign use case (WB 70%, ITS 15%, HA 15%)
 Set per-session $L_s^{req}, \beta_s, C_s, \delta_s, A_s^{req}$
end for
for each hexagonal sub-region h **do**
 Set objective: $\min \alpha \times \sum_s L_s + (1-\alpha) \times R_n^{used}$ (Equation (12))
 Add constraints: security (19), latency (14), bandwidth (15), resource (16), capacity (17)
 Run Gurobi; **if** feasible: encode optimal path as nix-vector, extend to DN via shortest path, store in F_{vector} ; **else**: mark session as dropped
end for
Return F_{vector}

Algorithm 3 NS3 internet stack initialization and traffic simulation

Input: M_{Adj} (from Algorithm 1), F_{vector} (from Algorithm 2)
Output: Per-session latency, drop rate, UPF utilization
 {—Network topology—}
 $nodesC \leftarrow \text{NodeContainer of } |M_{Adj}| \text{ nodes}$
for each (i, j) **with** $M_{Adj}[i, j] = 1$ **do**
 Install PointToPointHelper on (i, j) ; set delay $\leftarrow L_{i,j}$, DataRate $\leftarrow 100$ Gbps
end for
 {—Internet stack and routing—}
 Install custom Ipv4NixVectorHelper and InternetStackHelper on all nodes
 Assign IPv4 addresses to all point-to-point interfaces
 {—Applications and traffic—}
for each session i **in** F_{vector} **do**
 Install PacketSink on destination node
 Install OnOffHelperNixVector on source node; inject $nixVec \leftarrow F_{vector}[i]$; set β_s ,
 PacketSize = 256 B
 Start both applications at 1 s; stop at t_{stop}
end for
 {—Execution and metrics—}
 Simulator::Run() then Simulator::Destroy()
for each session s **do**
 if $L_s > L_s^{req}$ **or** session not completed **then**
 Increment drop counter
 end if
end for
Return average latency, drop rate, resource usage, average UPFs per session

6. Performance Evaluation

All experiments were carried out on a computer with 64 GB of RAM and a 12th generation Intel (R) Core (TM) processor. The proposed solution was developed using Python 3.10.12, Gurobi Optimizer Version 12.0.0 Build v12.0.0rc1, and C++, version 23 for the NS3 (ns-3.42) core module.

A direct numerical comparison with existing optimization models is not feasible because no prior work simultaneously addresses the three dimensions that define our problem: latency, PLMN operator cost, and security expressed as AC rules enforced by existing UPFs. While a direct comparison is infeasible, the computational performance of our model can be positioned against representative state-of-the-art models. To achieve this, we reduced the size of our network so that it more closely resembles the baseline models against which we are comparing (reduced to 150 nodes) and limited the number of PDU sessions to 200. As summarized in Table 3, our model achieves a competitive execution time.

Table 3. Execution time.

Work	Objective	Solution Method	Execution Time (s)
Stringer [49]	Latency vs. resource usage in SFC placement	MIP + heuristic	≈ 23 s
Leyva-Pupo et al. [18]	UPF placement reconfiguration	ILP (exact) + OST scheduler	≈ 7 s
DACR (ours)	Latency + PLMN cost, distributed AC rules	MILP	8.567 s

To nonetheless assess the relevance of our approach under directly comparable conditions, we therefore evaluate our model against a second model that we designed as a faithful representative of the security-agnostic, centralized-enforcement strategy implicit in the literature, in which all AC rules are enforced at the first nodes of the path rather than distributed along it. In this alternative model, Gurobi is employed to compute the shortest path for routing each PDU session. All other parameters remain unchanged. The following figures present the results of this comparison. We used the following performance metrics:

- Latency and percentage of resources consumed per number of PDU sessions generated for each use case;
- Rate of PDU session drop;
- Average number of processing UPFs per session.

For clarity, the model elaborated in Section 4 will be referred to as the DACR (Distributed Access Control Rules) model, while the second model used for comparison will be referred to as the comparison model.

We used the following number of rules as a standard:

- Web Browsing: A minimum of 15 basic rules and 10 complex rules.
- ITS: A minimum of 10 basic rules and five complex rules.
- Healthcare Applications: A minimum of 12 basic rules and three complex rules.

While Figure 2a expresses session loss as a percentage, Figure 2c presents the same outcome in absolute terms, namely, the number of PDU sessions that are successfully routed. The gap between the two models widens with the offered load. This confirms that the latency reduction achieved by distributing access control rules translates directly into higher effective network capacity.

Based on Figure 3 we conclude, for the web browsing use case, that our DACR model has moderately better performance than the comparison model: we observe a lower drop rate regardless of the number of PDU sessions. When a large number of PDU sessions is generated, the imposed latency threshold (140 ms) is consistently maintained, and the drop rate is 1.7% lower. The similarity in results can be attributed to the fact that the latency requirement is less stringent, which facilitates a greater number of potential routing options. Consequently, even under conditions of high network congestion, a viable route can still be identified.

By contrast, Figure 2 reveals a significant disparity between the models for the other two use cases. The DACR model demonstrates superior performance; while respecting the imposed latency threshold, it is capable of routing a greater number of PDU sessions. Furthermore, as the number of PDU sessions increases, the drop rate gap between the two models becomes more pronounced. Specifically, for 2,100,000 PDU sessions, the drop rates for the ITS use case are 12.54% for the DACR model and 23.49% for the comparison model, while for the health application use case, the drop rates are 16.58% and 25.26%, respectively.

Figure 3 offers a broader perspective by analyzing the overall generated traffic. It demonstrates that, overall, the DACR model consistently has a lower drop rate, regardless of the number of PDU sessions. For 700,000 sessions, the drop rates are 0.45% and 1.4% for the DACR model and comparison model, respectively, while for 2,100,000 sessions, they increase to 8.74% and 12.12%, respectively.

The strength of our DACR model lies primarily in its ability to manage sessions with strict latency requirements while maintaining performance for more flexible traffic flows. While the observed results for the web browsing use case are similar between the two models, the DACR model demonstrates superior performance in the other two use cases, which have more stringent latency requirements. Specifically, the DACR model outperforms the comparison model in terms of lower average latency and also, more

importantly, in terms of its ability to maintain a significantly lower and more reasonable drop rate.

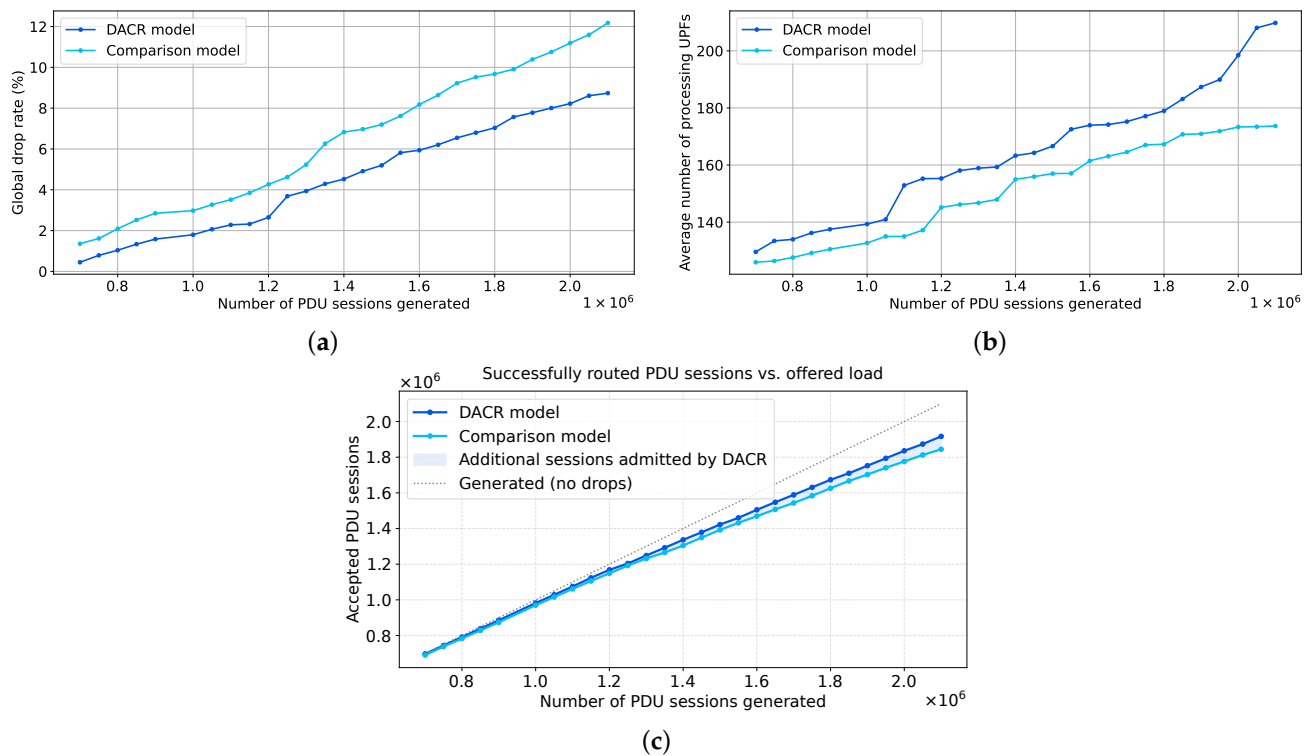


Figure 2. (a) Global drop rate per number of PDU sessions generated. (b) Average number of processing UPFs per session per number of PDU sessions generated. (c) Number of successfully routed PDU sessions as a function of the offered load.

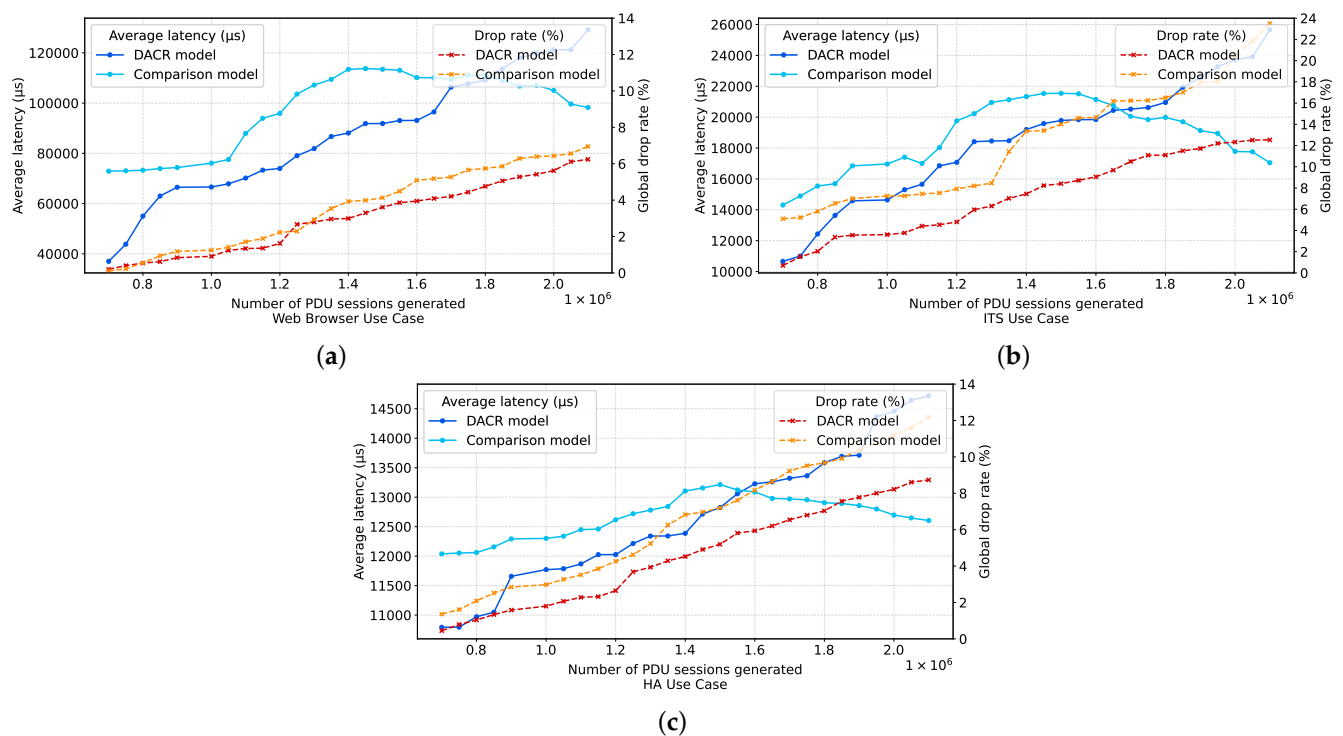


Figure 3. Average latency and drop rate per Number of PDU sessions generated per use case: (a) web browsing use case; (b) ITS use case; and (c) healthcare applications use case.

7. Conclusions

In this paper, we address the joint optimization of latency and resource consumption under security constraints in 5G networks. Security constraints are defined as a set of access control rules imposed on PDU sessions, with the AC rules distributed across all UPFs within the network. We investigate three use cases: web browsing, intelligent transport systems, and healthcare applications. This problem is formulated as a mixed integer linear programming problem aimed at minimizing round-trip latency and PLMN operator costs by optimizing the PDU session path while ensuring compliance with security and latency constraints. The commercial solver Gurobi was employed to determine the optimal solution, and a discrete event network simulator (NS3) was utilized to simulate the model in an environment that approximates real-world conditions.

Simulation results demonstrate the relevance of our model, particularly in comparison with a model without distributed AC rules. The effectiveness of the model is especially apparent for PDU sessions with stringent latency requirements. In these cases, when the network is congested with a high number of PDU sessions, the drop rate remains significantly lower. These results highlight the effectiveness of the model proposed in this article.

Several directions emerge for future research. A first direction concerns the design of a reconfiguration mechanism to reroute active PDU sessions, in order to adapt existing traffic to the arrival of new sessions with low latency requirements. A second direction is the optimization of access control rule placement across the UPFs, which we treated here as an operator-defined input rather than a decision variable.

A third direction concerns the applicability of our approach to upcoming 6G networks and, in particular, the scalability of the optimization framework. While this work focuses on the 5G user plane, the distributed AC rule enforcement proposed here aligns with the AI-native architecture that characterizes 6G [50], where security enforcement is expected to remain embedded within the user plane. The main limitation in our current approach lies not in the model itself but in its resolution method: as the optimization problem is NP-hard, computing the exact optimal solution becomes computationally intractable for the large-scale, dynamic, and real-time re-optimization scenarios that characterize 6G. The AI-native nature of 6G offers a natural answer to this challenge. Rather than solving the MILP problem exactly at each reconfiguration, learning-based methods can be trained to approximate the optimal PDU session path and AC rule distribution in near-constant time.

Author Contributions: Conceptualization, A.-G.C.; methodology, A.-G.C.; software, A.-G.C.; validation, J.V. and D.E.; formal analysis, A.-G.C.; writing—original draft preparation, A.-G.C.; writing—review and editing, J.V. and D.E.; supervision, D.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miuccio, L.; Panno, D.; Pisacane, P.; Riolo, S. A QoS-aware and channel-aware radio resource management framework for multi-numerology systems. *Comput. Commun.* **2022**, *191*, 299–314. [[CrossRef](#)]
2. Farooqi, A.M.; Alam, M.A.; Hassan, S.I.; Idrees, S.M. A fog computing model for VANET to reduce latency and delay using 5G network in smart city transportation. *Appl. Sci.* **2022**, *12*, 2083.

3. Jaaz, Z.A.; Ansari, M.D.; JosephNg, P.S.; Ghenni, H.M. Optimization technique based on cluster head selection algorithm for 5G-enabled IoMT smart healthcare framework for industry. *Paladyn J. Behav. Robot.* **2022**, *13*, 99–109.
4. Zhang, S.; Wang, Y.; Zhou, W. Towards secure 5G networks: A survey. *Comput. Netw.* **2019**, *162*, 106871.
5. Salahdine, F.; Han, T.; Zhang, N. Security in 5G and beyond: Recent advances and future challenges. *Secur. Priv.* **2023**, *6*, e271.
6. Cao, J.; Ma, M.; Li, H.; Ma, R.; Sun, Y.; Yu, P.; Xiong, L. A survey on security aspects for 3GPP 5G networks. *IEEE Commun. Surv. Tutor.* **2019**, *22*, 170–195.
7. Wu, T.; Guo, X.; Chen, Y.; Kumari, S.; Chen, C. Amassing the security: An enhanced authentication protocol for drone communications over 5G networks. *Drones* **2021**, *6*, 10. [\[CrossRef\]](#)
8. Rahman, A.U.; Mahmud, M.; Iqbal, T.; Saraireh, L.; Kholidy, H.; Gollapalli, M.; Ahmed, M.I.B. Network anomaly detection in 5G networks. *Math. Model. Eng. Probl.* **2022**, *9*, 397–404.
9. Gallenmüller, S.; Naab, J.; Adam, I.; Carle, G. 5G QoS: Impact of security functions on latency. In *NOMS 2020–2020 IEEE/IFIP Network Operations and Management Symposium*; IEEE: New York, NY, USA, 2020; pp. 1–9.
10. Parvez, I.; Rahmati, A.; Guvenc, I.; Sarwat, A.I.; Dai, H. A survey on low latency towards 5G: RAN, core network and caching solutions. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 3098–3130.
11. Bozorgchenani, A.; Zarakovitis, C.C.; Chien, S.F.; Ting, T.O.; Ni, Q.; Mallouli, W. Novel modeling and optimization for joint cybersecurity-vs-QoS intrusion detection mechanisms in 5G networks. *Comput. Netw.* **2023**, *237*, 110051.
12. Bozorgchenani, A.; Zarakovitis, C.C.; Chien, S.F.; Ni, Q.; Gouglidis, A.; Mallouli, W.; Lim, H.S. Intrusion response systems for the 5G networks and beyond: A new joint security-vs-QoS optimization approach. *IEEE Trans. Netw. Sci. Eng.* **2024**, *11*, 3039–3052. [\[CrossRef\]](#)
13. Fadlullah, Z.M.; Wei, C.; Shi, Z.; Kato, N. GT-QoSec: A game-theoretic joint optimization of QoS and security for differentiated services in next generation heterogeneous networks. *IEEE Trans. Wirel. Commun.* **2016**, *16*, 1037–1050.
14. Iordache-Sica, M.M.; Anagnostopoulos, C.; Pezaros, D.P. Towards QoS-aware provisioning of chained virtual security services in edge networks. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*; IEEE: New York, NY, USA, 2021; pp. 178–186.
15. Bagaa, M.; Taleb, T.; Bernabe, J.B.; Skarmeta, A. QoS and resource-aware security orchestration and life cycle management. *IEEE Trans. Mob. Comput.* **2020**, *21*, 2978–2993.
16. Doriguzzi-Corin, R.; Scott-Hayward, S.; Siracusa, D.; Savi, M.; Salvadori, E. Dynamic and application-aware provisioning of chained virtual security network functions. *IEEE Trans. Netw. Serv. Manag.* **2019**, *17*, 294–307.
17. Miskeen, G.; Kouvatsos, D.D.; Akhlaq, M. Performance and security trade-off for routers in high speed networks. In *26th UK Performance Engineering Workshop (UKPEW 2010)*; University of Bradford: Bradford, UK, 2010; pp. 119–128.
18. Leyva-Pupo, I.; Cervelló-Pastor, C.; Anagnostopoulos, C.; Pezaros, D.P. Dynamic UPF placement and chaining reconfiguration in 5G networks. *Comput. Netw.* **2022**, *215*, 109200.
19. Chen, S.; Chen, J.; Li, H. Joint optimization of UPF placement and traffic routing for 5G core network user plane. *Comput. Commun.* **2024**, *216*, 86–94.
20. Li, Y.; Ma, X.; Xu, M.; Zhou, A.; Sun, Q.; Zhang, N.; Wang, S. Joint placement of UPF and edge server for 6G network. *IEEE Internet Things J.* **2021**, *8*, 16370–16378. [\[CrossRef\]](#)
21. Shameli-Sendi, A.; Jarraya, Y.; Pourzandi, M.; Cheriet, M. Efficient provisioning of security service function chaining using network security defense patterns. *IEEE Trans. Serv. Comput.* **2019**, *12*, 534–549.
22. Wang, B.; Li, J.; Cao, S.; Guler, E.; Zheng, D. Security-aware service function chaining and embedding with asymmetric dedicated protection. *IEEE Access* **2024**, *12*, 53944–53957.
23. Tamim, I.; Jammal, M.; Hawilo, H.; Shami, A. Introducing virtual security functions into latency-aware placement for NFV applications. In *ICC 2020–2020 IEEE International Conference on Communications (ICC)*; IEEE: New York, NY, USA, 2020; pp. 1–7.
24. Dwiardhika, D.; Tachibana, T. Virtual network embedding based on security level with VNF placement. *Secur. Commun. Netw.* **2019**, *2019*, 5640134. [\[CrossRef\]](#)
25. Wang, C.; Zheng, D.; Liu, X.; Tang, W.; Xu, H.; Cao, X. Towards cost optimization in security-aware service function chaining and embedding over multi-vendor edge networks. *Comput. Netw.* **2025**, *257*, 111002. [\[CrossRef\]](#)
26. Hoang, T.T.; Pham, L.M.; Nguyen, H.S. LAVP: A latency-aware virtual network function placement strategy for service function chain in network function virtualization. *IEEE Access* **2025**, *13*, 151327–151337. [\[CrossRef\]](#)
27. 3rd Generation Partnership Project. System architecture for the 5G System (5GS) (Technical Specification No. TS 23.501 V18.6.0). 3GPP. 2024. Available online: https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/ (accessed on 15 June 2026).
28. 3rd Generation Partnership Project. Procedures for the 5G System (5GS) (Technical Specification No. TS 23.502 V18.6.0). 3GPP. 2024. Available online: https://www.3gpp.org/ftp/Specs/archive/23_series/23.502/ (accessed on 15 June 2026).
29. 3rd Generation Partnership Project. Interface between the Control Plane and the User Plane nodes (Technical Specification No. TS 29.244 V18.6.0). 3GPP. 2024. Available online: https://www.3gpp.org/ftp/Specs/archive/29_series/29.244/ (accessed on 15 June 2026).

30. 3rd Generation Partnership Project. Security architecture and procedures for 5G System (Technical Specification No. TS 33.501 V18.9.0). 3GPP. 2024. Available online: https://www.3gpp.org/ftp/Specs/archive/33_series/33.501/ (accessed on 15 June 2026).
31. Michaelides, S.; Mucke, J.; Henze, M. Assessing the latency of network layer security in 5G networks. In *18th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec 2025)*; ACM Digital Library: New York, NY, USA, 2025; pp. 262–267.
32. Lucas-Estañ, M.C.; Coll-Perales, B.; Shimizu, T.; Gozalvez, J.; Higuchi, T.; Avedisov, S.; Sepulcre, M. An analytical latency model and evaluation of the capacity of 5G NR to support V2X services using V2N2V communications. *IEEE Trans. Veh. Technol.* **2022**, *72*, 2293–2306.
33. Garey, M.R.; Johnson, D.S.; Stockmeyer, L. Some simplified NP-complete problems. In *Sixth Annual ACM Symposium on Theory of Computing*; ACM Digital Library: New York, NY, USA; pp. 47–63.
34. Sallam, G.; Gupta, G.R.; Li, B.; Ji, B. Shortest Path and Maximum Flow Problems Under Service Function Chaining Constraints [Preprint]. *arXiv* **2018**, arXiv:1801.05795. Available online: <https://arxiv.org/abs/1801.05795> (accessed on 15 June 2026).
35. Sallam, G.; Ji, B. Joint placement and allocation of VNF nodes with budget and capacity constraints. *IEEE/ACM Trans. Netw.* **2021**, *29*, 1238–1251. [[CrossRef](#)]
36. Ma, C.S.; Miller, R.H. MILP optimal path planning for real-time applications. In *2006 American Control Conference*; IEEE: New York, NY, USA, 2006; pp. 1–6.
37. Jain, V.; Grossmann, I.E. Algorithms for hybrid MILP/CP models for a class of optimization problems. *Inf. J. Comput.* **2001**, *13*, 258–276.
38. 3rd Generation Partnership Project. Study on 3D Channel Model for LTE (Technical Report No. TR 36.873 V12.7.0). 3GPP. 2017. Available online: https://www.3gpp.org/ftp/Specs/archive/36_series/36.873/ (accessed on 15 June 2026).
39. Shabiluh, A.; John, U.K.; Esenogho, E.; Sadiq, B.O.; John, A.; Dintwa, E.; Sheikh, S.; Sajid, M.; Matlotse, E. Throughput and latency performance evaluation of an optical fiber network. *J. Adv. Inf. Technol.* **2025**, *16*, 71–80. [[CrossRef](#)]
40. Fettweis, G.; Alamouti, S. 5G: Personal mobile internet beyond what cellular did to telephony. *IEEE Commun. Mag.* **2014**, *52*, 140–145. [[CrossRef](#)]
41. Moubayed, A.; Shami, A.; Heidari, P.; Larabi, A.; Brunner, R. Edge-enabled V2X service placement for intelligent transportation systems. *IEEE Trans. Mob. Comput.* **2020**, *20*, 1380–1392.
42. Lassagne, A. Retour Sensoriel du Toucher Dans les Environnements Virtuels. Ph.D. Thesis, École Nationale Supérieure d’arts et métiers–ENSAM, Paris, France, 2018.
43. Ahmadi, S. Next generation AI-based firewalls: A comparative study. *Int. J. Comput.* **2023**, *49*, 245–262.
44. Cereia, M.; Bertolotti, I.C.; Durante, L.; Valenzano, A. Latency evaluation of a firewall for industrial networks based on the Tofino industrial security solution. In *2014 IEEE Emerging Technology and Factory Automation (ETFA)*; IEEE: New York, NY, USA, 2014; pp. 1–8.
45. Lekidis, A.; Anastasiadis, A.G.; Vokas, G.A. Anomaly detection mechanisms for in-vehicle and V2X systems. In *9th ACM Cyber-Physical System Security Workshop (CPSS ’24)*; ACM Digital Library: New York, NY, USA, 2024; pp. 1–8.
46. Sedjelmaci, H.; Senouci, S.M.; Ansari, N. Security of 5G-V2X: Technologies, standardization and research directions. *IEEE Netw.* **2019**, *33*, 62–69. [[CrossRef](#)]
47. Alves, F.; Ribeiro, R.; Petiz, M.; Abbasi, A.; Carvalho, P.; Faia, R.; Faia, P.; Zaia, R.; Rodrigues, R. Virtual power plant optimization service-benchmark of solvers. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*; IEEE: New York, NY, USA, 2024; pp. 279–287.
48. Jablonský, J. Benchmarks for current linear and mixed integer optimization solvers. *Acta Univ. Agric. Silvic. Mendel. Brun.* **2015**, *63*, 1923–1928.
49. Kulkarni, S.G.; Arumaithurai, M.; Ramakrishnan, K.K.; Fu, X. Stringer: Balancing latency and resource usage in service function chain provisioning [Preprint]. *IEEE Internet Comput.* **2016**, *20*, 22–31.
50. Porambage, P.; Gür, G.; Moya Osorio, D.P.; Liyanage, M.; Gurtov, A.; Ylianttila, M. The roadmap to 6G security and privacy. *IEEE Open J. Commun. Soc.* **2021**, *2*, 1094–1122. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.