IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 16, 2020, accepted September 1, 2020, date of publication September 8, 2020,
date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022741

When Agile Security Meets 5G

GLAUCIO H. S. CARVALHO[1], **ISAAC WOUNGANG**[1], **(Senior Member, IEEE), ALAGAN ANPALAGAN**[2], **(Senior Member, IEEE), AND ISSA TRAORE**[3]

[1]Department of Computer Science, Ryerson University, Toronto, ON M5B 2K3, Canada
[2]Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada
[3]Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 2Y2, Canada

Corresponding author: Glaucio H. S. Carvalho (glauciohscarvalho@gmail.com)

This work was partially supported by a grant from the National Science and Engineering Research Council of Canada (NSERC) held by the second author (REF#: RGPIN 2017 04423) and in part by Discovery Grant held by the third author.

ABSTRACT 5G is a critical infrastructure that will connect the whole society and bridge other critical infrastructure systems. Thus, cybersecurity emerges as crucial tenet within the 5G pathway. In this article, we discuss the concept of agile security within a 5G infrastructure taking into account two of its major technologies: Mobile Edge Computing (MEC) and Network Functions Virtualization (NFV). In this sense, we first discuss the 5G-driven MEC deployment from a NFV perspective. Secondly, we present the concept of agile security and how it can be embedded in the daily activities of mobile network operators (MNOs). Thirdly, we discuss risk management as a key element of the agile security framework. To illustrate its application, we propose the design of an agile security risk-aware edge server mechanism for 5G driven MEC deployment, which uses multiple thresholds and a load-balancing security control to mitigate the risks of resource exhaustion and violation of the service level agreement (SLA) faced by the edge servers while taking advantage of multiple cloud layers to increase the degree of availability and dependability of the system. Numerical results show that the proposed mechanism is able to keep the risk at lower levels.

INDEX TERMS 5G security, agile security, cloud security, security risk management.

I. INTRODUCTION

5G-driven Mobile Edge Computing (MEC) will trigger the development of services and applications that will empower all societal sectors through the use of a densified, pervasive, ultra-reliable, sustainable, and performant computing and communications infrastructure. This advanced network architecture will underpin the existence of smart cities, smart grids, smart factories, to name a few emergent critical infrastructure. Due to its criticality and key role, a successful realization of the 5G-driven MEC will call for an agile security approach.

Agile security is a process of rapidly responding to the changes in a highly volatile cyberspace by leveraging the appropriate resources and system controls to shield the assets against unauthorized, unintended, and malicious actions. To perform such operation with high degree of efficiency, mobile network operators (MNOs) need to continuously, ubiquitously, and proactively monitor their infrastructure in order to quickly detect and instantly react to defeat the nefarious activities. Since the deployment of a large number of physical security devices such as intrusion detection systems (IDSs) and firewalls across the densified 5G infrastructure might be cost- and performance-prohibitive, a solution should be sought in order to realize an agile security approach in a 5G-driven MEC system.

In the purview of network management and optimization, the concept of network functions virtualization (NFV) arises as a compelling solution for the problem of agile security. Indeed, under this framework, networking functionalities are softwarized to become Virtualized Network Functions (VNFs) [1], [2]. The VNFs can be wrapped within a virtual machine (VM), instantiated, monitored, and chained with others to provide a rich set of network services in a flexible way. In terms of security, doing so will pave the way for the creation of a more affordable, customizable, and innovative solutions with the use of security service chaining (SSC) for instance. Using the speed, versatility, and manageability of virtualized technologies, MNOs are able to meet the requirements of the agile security approach where security services such as virtual IDS and firewalls can be scaled out by virtual load balancers to mitigate the attacks anywhere in the network.

From an operational perspective, security services will dispute the virtual resources with mobile cloud

The associate editor coordinating the review of this manuscript and approving it for publication was Hongbin Chen.

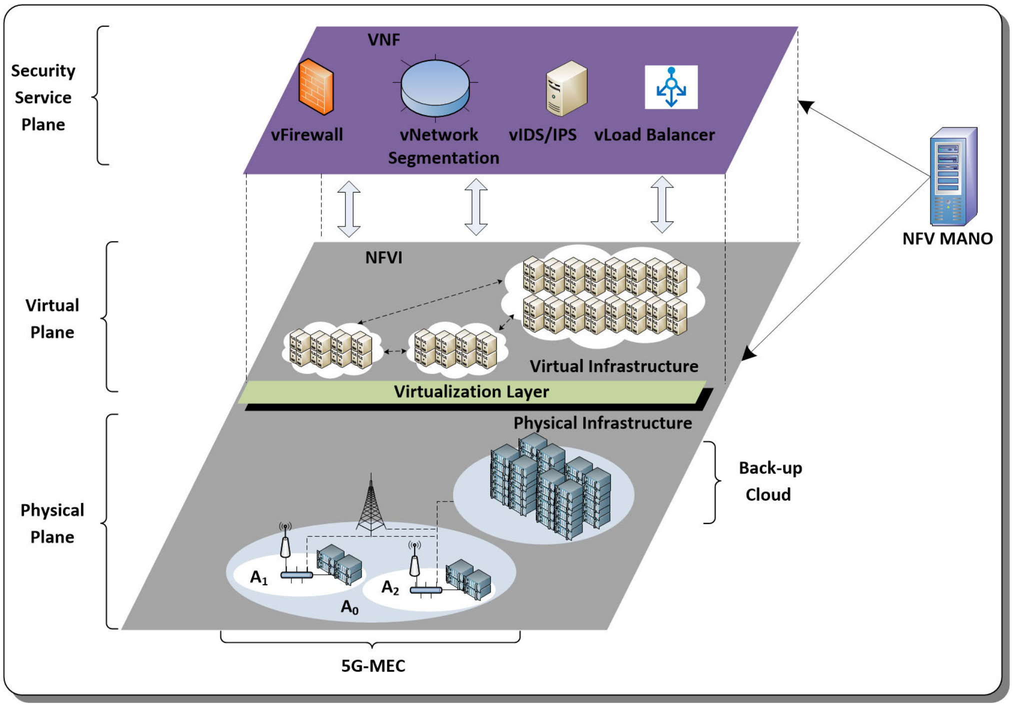

FIGURE 1. 5G-driven MEC enabled by NFV.

computing (MCC) applications and services. While in a typical cloud data center, the abundance of virtual resources can ease the provision of the SSC, the same cannot hold about the MEC. Indeed, at the edge of the network, the inherent resource-constrained nature of the cloud system might become a barrier for MNOs to secure their assets in an agile manner if the cloud orchestrator does not intelligently provision and release the virtual resources at various cloud layers in a cooperative manner [3]. Thus, the success of the agile security in a 5G-driven MEC deployment is tightly coupled to the problem of allocation of communications and computing resources.

Despite the fact that 5G, MEC, and NFV have gained a lot of momentum in the literature, to the best of our knowledge, a research work that integrates all them holistically under the umbrella of agile security is unprecedented. In order to bridge this gap in literature, this article sheds light on the operationalization aspects of the agile security considering the amalgamation of 5G, MEC, and NFV.

To showcase how these technologies can be integrated synergistically, we propose a security-aware risk aware edge server orchestrator that takes advantage of multiple cloud layers to augment the capacity and the degree of availability and dependability of the edge servers while mitigating the risks of resource exhaustion due to a denial of service (DoS) attack and flash workload in addition to the Service Level Agreement (SLA) violation, which is crucial to ensure the proper QoS and QoE of resource-hungry applications that require to be processed at the nearest MEC site to meet their low latency requirement. To mitigate the aforementioned risks, the proposed orchestrator applies two security controls, namely: load balancing mechanism and multiple thresholds, which work holistically to keep the security risks under acceptable levels. The load balancing mechanism is designed to select the appropriate edge server to accommodate an incoming service request and to migrate the virtual machines (VMs) between the cloud layers taking into account the aforementioned security and performance requirements while the thresholds are designed to limit the number of untrusted users into the system. A continuous time Markov chain (CTMC) model is presented and risk figures are derived using the CTMC's steady state distribution.

The rest of this article is organized as follows. Section II presents an overview of 5G, MEC, and NFV and the need for an efficient risk-based resource management. Section III presents the framework of agile security and our proposed orchestration process that will make it viable. Section IV describes the role played by risk management in agile security. Section V presents a review of the literature on security risk-aware resource allocation and describes the design of the proposed agile security risk-aware edge server mechanism within the context of a 5G-driven MEC deployment. Numerical results, which are presented in Section VI, show that the proposed scheme is effective in keeping the security risk under lower levels. Finally, concluding remarks and future research directions are presented in Section VII.

II. 5G-DRIVEN MEC NETWORK ARCHITECTURE ENABLED BY NFV

The NFV architecture is formed by the NFV infrastructure (NFVI), the Virtual Network Functions (VNFs) and the NFV Management and Orchestration (NFV MANO) [1].

Fig. 1 illustrates the realization of our proposed 5G-driven MEC deployment enabled by the NFV technology to support the agile security. The network architecture is explained as follows:

- NFVI defines the underlying physical and the virtual infrastructure that underpins the provisioning of the softwarized security services. Considering the 5G-MEC components, the NFVI features the edge servers, the network connections, and the back-up cloud data center as major physical components. The back-up cloud defines a resource-rich cloud data center deployed by the MNOs that can be used for example to carry out the bursty workload from the edge servers. The NFVI also defines the infrastructure as a service (IaaS) that results from the virtualization of the MEC and back-up cloud. Consequently, considering the proposed architecture, the NFVI comprises of the physical plane and virtual plane presented in Fig. 1.
- VNFs are the implementation of the security services such as virtual firewalls, virtual network segmentation, virtual IDS/IPS, and virtual load balancers, to name a few, that run on top of the Infrastructure-as-a-Service (IaaS). Within our proposed architecture, these security services belong to the security service plane. The definition of a VNF unleashes the creation of innovative services that can be tailored to meet a specific security demand. For instance, a VNF can execute a specific security service such as firewalling, or a combination of VNFs and network resources can be chained to create a virtual segmented network with firewalls in which each network segment accommodates the traffic of a specific service class while isolating the communication between different service classes.
- NFV MANO is responsible for the operationalization of the entire physical infrastructure, the virtual infrastructure, and the VNFs which includes the creation, deployment, and termination of security services as well as the dynamic allocation of physical/virtual resources that will be used by the VNFs. As a result, the NFV MANO holistically orchestrates the resources pertaining to the physical plane, the virtual plane, and the security service plane to guarantee a high availability, reliability, and performance of the security services.

III. AGILE SECURITY

In a dynamic threat landscape as the one faced by the MNOs, the provisioning of agile security is instrumental to ensure a less riskier operation. As a process, agile security is featured by a continuous, ubiquitous, and proactive monitoring of the system infrastructure in such a way that the system can instantly and autonomously react to mitigate an attack when it is detected or when its risk reaches a pre-specified level.

To better understand the challenges of agile security in a 5G-driven MEC deployment, we present as follows a discussion considering its three features, namely: continuous monitoring, ubiquitous monitoring, and proactive monitoring.

A. CONTINUOUS MONITORING

To continuously monitor their infrastructure, the MNOs must have the ability to uninterruptedly verify the status of the computing, storage, and network components in the physical and virtual planes as well as the VNFs in the security service plane running on their multiple cloud systems. For an effective 24/7 monitoring to take place, the NFV MANO must have services in place that navigate seamlessly throughout the three planes and provide an updated status of the physical servers, the hypervisors, the virtual servers, the network connections, and the interactions between planes. It also should be able to take the right action when a malicious intent is found or a threshold is reached.

The success of a continuous monitoring depends on the pervasive capillarity of security services, which in turns implies the ability to reach every network segment of the virtual and physical ultra-dense 5G infrastructure. Additionally, the health of VNFs that deliver security services should be monitored and fault-tolerance procedures, which will ensure an interrupted service delivery in case of a failure, should be operational. Furthermore, a deep inspection of the virtualized infrastructure is paramount for securing the three planes holistically. In this respect, hypervisor introspection procedures that parse the running VMs can be applied to secure the operations of the VMs, the corresponding hypervisor, and consequently the corresponding physical server [1], [4]. By monitoring the communication services, storage services, and computing services of a VM, introspection techniques provide the hypervisors with the ability to detect abnormal activities that is critical for the success of a continuous monitoring.

B. UBIQUITOUS MONITORING

Given the comprehensiveness and the densified nature of 5G systems, ubiquitous monitoring will be possible only if the MNOs have enough networking and computing resources in every network segment to support the execution of security services, the system operation, and the MCC services simultaneously. Nonetheless, due to the proximity to end users, openness and freedom of wireless communications, intricacies of cloud configurations, and inherent resource-constrained nature, it is likely that hackers are going to primarily focus on the edge servers. In this sense, the MNOs can respond proactively by creating security policies and mechanisms that intelligently exploit the presence of multiple cloud layers and wireless connections to cooperatively mitigate the risk across the entire infrastructure. For instance, secure VM migration techniques can be triggered to transfer latency-insensitive applications to the back-up cloud to free up enough capacity at the edge network in such a way that a SSC comprised of virtual IDP/IPS and virtual firewalls could be leveraged.

Since ubiquitous monitoring will require the instantiation, monitoring, and possibility the chaining of VNFs in an on-the-fly manner, it is paramount to prevent the tampering of

VNFs from happening while its image is transferred to the edge server. The detection of a malfunctioning or corrupted security service, which is supposed to safeguard the system, might be a difficult undertaking, which can potentially wreak havoc the entire MEC site while creating a security hole for hackers to exploit to. The use of a private key infrastructure to verify the integrity of the secure VNFs prior to their launching arises as a viable solution to this issue [4].

It is worth noting that a failure in the provisioning of ubiquitous monitoring might put extra pressure on the continuous monitoring which may take longer to detect a malicious activity (or its effect) that occurs in an unmonitored network segment. This delayed detection might be just the time that hackers need to adversely affect the system or exfiltrate the user data.

C. PROACTIVE MONITORING

Proactive monitoring refers to the ability of anticipating the threats and mitigating their risks. To this end, the MNOs should offensively attack their own infrastructures using ethical hacking and implement the respective system controls to avoid the erosion of the system security.

Several benefits can be harvested from a proactive monitoring. Firstly, the MNOs will raise their awareness on their own architecture and exploitable vulnerabilities. Secondly, by discovering and fixing the security flaws in the earlier stages, they can reduce the system exposition, which will give less time for hackers to take advantage of the security holes. Thirdly, they can minimize the uncertainty of handling an attack by anticipating potential incidents and devising a plan that will save them from experiencing financial losses due to data theft, system unavailability, damage to brand reputation, to name a few. Last but not least, proactive monitoring is a practice that will incrementally enhance the continuous and ubiquitous monitoring. For instance, the red team will systematically attack the building blocks of the physical plane, the virtual plane, and the security service plane while the blue team will defend against these attacks by leveraging the omnipresence and orchestration ability of the NFV MANO. In this respect, the blue team can customize innovative security services for a specific attack such as creating a chain of detection, prevention, and firewalling services on-demand or apply security controls such as disconnecting virtual network segments, or replacing an existing compromised physical server while distributing its operational and healthy VMs across multiple cloud layers. As a consequence, the outcome of every red-blue team simulation is a more effective defense system.

IV. SECURITY RISK-AWARE ORCHESTRATION

To be agile, a security framework should be designed around the principle of risk management. The reason for a security risk-aware approach is the fact that the orchestration of the cloud resources and corresponding decision making process should take into consideration the security risk inflation and deflation while managing the virtual resources, physical resources, and services to accomplish a more performant and reliable operation. The specification of the security risk revolves around the definition of the asset. In a 5G-driven MEC network, an asset is a valuable object (i.e. data, network element, hardware, software, process or any other component that supports information-related activities in the considered environment) that must be safeguarded in order for the MNO to keep and gain competitiveness and market share.

Assets might be susceptible to a number of threats that endanger a safe service provisioning. If a vulnerability that is a security flaw or weaknesses exists, then a threat can exploit it to harm and reduce the value of the asset. In this respect, the security risk can be computed as the likelihood or probability of such exploitation to occur times its consequence or loss impact. Since a successful exploit might undermine the MNO's operation and consequently its reputation, the specification of a risk management framework that mitigates the potential breaches of confidentiality, integrity, and availability (CIA) of the services while protecting the data pertaining to the clients and the MNO from attacks, becomes imperative for a sustainable and profitable operation.

A. RISK IDENTIFICATION

As stated earlier, there is no risk without an asset. By the same token, without a vulnerability, a threat cannot endanger an asset. Bearing these principles in mind, the risk identification phase aims at creating an inventory of the MNO's assets, vulnerabilities, and threats. In a 5G-driven MEC network like the one in Fig. 1, assets are defined by the components of the NFVI physical and virtual planes, the VNFs in the form of security services, and the NFV MANO. Taking into account the myriad of systems from the physical layer to the application layer, a full characterization of the threat landscape and the way that threats can potentially exploit an existing vulnerability might become an intricate task. For example, considering the Openstack and KVM hypervisors, which are fundamental 5G infrastructure systems [5], the Common Vulnerabilities and Exposures (CVE) website has listed for these technologies a number of 271 and 157 publicly known cybersecurity vulnerabilities, respectively [6]. Moving to the application layer, Google Chrome and Maps, which are popular Web technologies, have a total of 2366 and 21 CVE entries while Twitter and Instagram account for 60 and 18 CVE entries, respectively [6]. In addition to taking advantage of known vulnerabilities, adversaries still can perform social engineering attacks to penetrate the system or to spread a malware infection. In fact, the Anti-Phishing Working Group (APWG) Q1 2020 report has shown that SaaS/Webmail represents 35% of the most targeted sectors for phishing attacks [7].

B. RISK ASSESSMENT

With a thorough characterization of the assets, threats, and known vulnerabilities pertaining to the MNO, the next step consists of quantitatively or qualitatively assigning a score to the risk. A score corresponds to a risk level that collectively enables the creation of a ranking of risks, which will ultimately be used to guide the agile security decision making process.

In a quantitative risk assessment approach, there is a need to derive the likelihood or probability that a specific vulnerability will be successfully exploited. Recall that an agile security framework requires a rapid risk-centric response to the changes in cyberspace. To accomplish such a task with a high degree of efficiency, the infrastructure should be continuously and ubiquitously monitored to trigger the application of adequate action as discussed in Section III. Thus, to accurately compute the probability of successful exploits across the entire infrastructure and services, the risk model must satisfy these requirements. Initiatives such as those reported in [8] in which a OPTIMIS-oriented risk assessor uses measurable data from a cloud infrastructure to compute the likelihood of events and the corresponding risk might pave the way for the application of an agile security framework on a 5G-MEC driven infrastructure.

To complete the security risk calculation, an analysis of the loss impact must be put forward. This step is featured by an educated choice for the loss impact. Traditionally, scales are used to quantify or qualify it. Instances of impact loss scales are: low, medium, and high [8] as well as very low (1), low (2), medium (3), high (4), and very high (5) [9].

C. RISK CONTROL

The last phase of a risk-aware operation defines the risk-handling posture adopted by the MNO that might be the *risk avoidance* where the MNO decides to circumvent the business activities that unfold the risk, the *risk acceptance* where the MNO decides to accept the exploit and its impact on the system; the *risk transference* where the MNO decides to shift to a third party, traditionally an insurance company, part or the total cost that the MNO will incur if a successful exploitation takes place, and the *risk mitigation* where the MNO decides to reduce the harmful impact of the risk by applying security controls and countermeasures that will keep the risk within the MNO's risk appetite [10].

In Section V, we present a stochastic risk-aware edge server orchestration that applies a risk mitigation strategy to handle the operational risks that the edge servers might be exposed to in a 5G-driven MEC deployment. The proposed risk-aware agile security method makes use of the loss impact scale introduced in [9].

V. SECURITY RISK-AWARE EDGE SERVER ORCHESTRATION

A. RELATED WORKS

Virtual machine (VM) allocation for cloudified infrastructure has been an intensive field of research. However, most of breakthroughs have neglected the role played by the security awareness and the risk awareness when it comes to the admission and placement of the task execution and services. In fact, when it comes to security risk-awareness, there is a scarcity of works in the literature. To enlarge our review of the literature, we address MEC, mobile cloud computing (MCC), and cloud computing.

Djemame *et al.* [8] proposed a OPTIMIS-aware risk assessor that collects measurable data from a cloud infrastructure to quantify the risk level. The cloud optimizer, which works continuously, takes into consideration the events of DoS, flash network traffic, hardware failure, and SLA violations and displays their risk levels. Due to the resource sharing nature of multi-tenancy, co-resident attacks become a serious threat to virtualized cloud infrastructure. Considering the stable and fluctuating features of a workload, Chhetri *et al.* [11] proposed a risk- and cost-aware resource allocation for cloud applications that seeks to minimize cloud costs while mitigating the resource revocation risks. Experimental results showed that transient resources can be leveraged to accomplish such objective.

In order to minimize the risk of data leakage among tenants, Almutairi *et al.* [12] proposed risk-aware virtual resource assignment mechanism targeting Software-as-a-Service (SaaS) cloud service providers (CSP). The authors proposed two cost functions (RASP-MAX and RASP-PAR) and different heuristic models (best fit heuristic, single move heuristic, and multi move heuristic) to solve the risk aware VM assignment problem. Considering metrics such as efficiency and coverage, Han *et al.* [13] proposed VM allocation policy to mitigate the risk of co-residence. The proposed scheme also took into account power consumption and workload balance as optimization goals. Similarly, the work in [14] proposed a security-aware VM placement mechanism to proactively avoid co-residency between malicious and benign VMs. Using a conflict index as a risk indicator, Miao *et al.* applied VM migration to protect benign users. Considering the learnings from [13] and [14], Han *et al.* [13] proposed a co-resident threat defense system, which comprises of security-aware VM management mechanisms and threat score mechanism, that mitigated the risk of attack while leading to balanced workload with little impact on power consumption. A commonality among [13], [14], and [15] is the use of CloudSim as the testing tool.

Within the purview of MEC and MCC, there has been a few initiatives for security- and risk-centric managers to support the orchestration of the cloud resources. A optimized allocation of the virtual resources to ensure performant and secure execution stand out as the major contributions behind the works [16], [17], and [18], where the differences lie in the fact that [16] applies to a system with multiple cloud layers while the others emphasized a standalone wireless and cloud system. Additionally, the authors provided a detailed cost function that can be used by MNOs to estimate the economical benefits of providing security services. Security risk takes the form of an user taxonomy in [19], where Raei *et al.* classified the users requests into three categories, namely: low risk, high risk, and critical risk. In this respect, the low risk category specifies applications that can be processed without cryptography and privacy concerns. On the

TABLE 1. Summary of related works.

Reference	Security and Risk Figures	MEC/Fog	Methodology
[9]	Resource exhaustion, hardware failure, and SLA violations	No	Experimentation
[12]	Resource revocation risk	No	Experimentation
[13]	Data leakage	No	Graph Theory and heuristic
[14]	Reduce the risk of co-resident attack	No	Simulation
[15]	Reduce the risk of co-resident attack	No	Simulation
[16]	Reduce the risk of co-resident attack	No	Simulation
[17]	Virtual resources to protect the task execution	Yes	Semi-Markov Decision Process
[18]	Virtual resources to protect the task execution	No	Semi-Markov Decision Process
[19]	Virtual resources to protect the task execution	No	Semi-Markov Decision Process
[20]	Virtual resources to protect the task execution	No	Markov reward model and simulated annealing
[21]	Security overhead to protect the task execution	Yes	Genetic Algorithm
[22]	Security overhead to protect the task execution	Yes	Deep Reinforcement Learning
[23]	Computation and communication uncertainties	Yes	Game Theory
[24]	Risk-neutral user, risk-averse user, risk-seeking user	Yes	Simulation
[25]	IDS at the edge of the network	Yes	Stochastic Differential Equation
[26]	Service failure	No	Graph Theory
[27]	Service and server failure	No	System Optimization and Heuristics
Our work	Resource exhaustion, user taxonomy, SLA violation (latency-awareness)	Yes	Markov Chain and Queueing theory

other hand, the high risk category requires multiple VMs to safeguard the execution of the offloaded task while the critical risk category demands an exclusive physical server to isolate the application. Considering the overhead that is needed to execute security services, Huang *et al.* proposed SEECO [20] and SCACO [21] mechanisms for secure task offloading in MEC. When it comes to the optimization problem, SEECO takes into account energy and security as design requirements while the SCACO additionally copes with processing delay. The papers are also dissimilar in the way that the authors solve the optimization problem. In this respect, SEECO is formulated as a genetic algorithm while SCAC is solved using deep reinforcement learning.

Taking the computation and communication uncertainties at each MEC server as the risk figure, Apostolopoulos *et al.* [22] formulated the problem of task offloading as a non-cooperative game among the users and solved it by means of the pure Nash Equilibrium. A risk-centric broker mechanism is proposed by Iyer *et al.* [23] in which the decision making problem is formulated to decide about the computing task placement between a cloud data center or a fog data center taking into account the user risk profile, price, and reputation of the cloud and fog service provider. With the goal of supporting security services at the edge of the network, Hui *et al.* [24] proposed a resource allocation scheme to efficiently share the edge resources between data processing services and security services, more specifically, intrusion detection system (IDS).

Service survivability in the presence of failure is addressed in [25] and [26] from a perspective of a NFV deployment. In [25], Kanizo *et al.* proposed to use a VNF to implement and deploy backup schemes for network functions that ensure high levels of survivability while reducing resource consumption. He *et al.* [26] proposed a backup resource allocation model where both network services and backup servers fail. Given the limited capacity of backup and the fact the network functions differ in terms of importance, the authors proposed to backup network functions based on their importance.

Table 1 summarizes the related works and shows how our proposal stands out in the literature. Firstly, like [8] it copes with resource exhaustion and SLA violation. In our case; however, a breach to SLA leads to an increased latency that might degrade the QoS and QoE inasmuch as we are dealing with MEC while [8] is not. Similar to [16]–[19], our work applies a user taxonomy to raise the awareness about the risk that a service request is posing to the system. However, we classify users rather than the applications. Additionally, the proposed algorithm takes advantage of multiple cloud layers to augment the capacity and the degree of availability and dependability of the edge servers as our previous work [16]. However, the current proposition specifies an implementation-friendly mechanism that can be integrated in a production system more effectively than an optimal controller that requires a sophisticate architecture. Furthermore, this work addresses the security risks at physical and virtual levels rather than a cloud level as [16]. The practicality of proposed orchestrator relies on the use of a load balancing mechanism and multiple thresholds to efficiently orchestrate the services requests while satisfying the security and latency requirements. The load balancing mechanism is proposed to choose the appropriate edge server to host the offloaded application and migrate VMs between the cloud layers taking into account the aforementioned requirements while the thresholds are designed to limit the number of untrusted users into the system.

B. PROPOSED MECHANISM

The proposed security risk-aware edge server orchestrator is designed to work in an mobile cloud computing (MCC) environment such as the one presented in Fig. 1, which is featured by a MEC and a back-up cloud. Under this assumption, it aims at mitigating the risk of resource exhaustion caused by a DoS attack (malicious activity) or a sudden spike in the traffic load (non-malicious activity) as well as the risk of SLA violation caused by a growth in latency due to the placement of the offloaded task in the remote back-up cloud. To mitigate the aforementioned risks, the proposed orchestrator applies two security controls, namely: load balancing mechanism and multiple thresholds, which work synergistically to keep the security risks under acceptable levels.

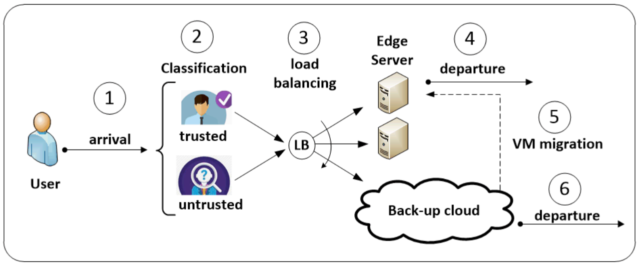

FIGURE 2. **Proposed risk-aware orchestrator operation.**

Fig. 2 depicts the proposed risk-aware edge server orchestrator from an operational standpoint. As we can see, upon an user arrival ① the system classifies the user into trusted and untrusted ones ②. This is necessary since the decision making process regarding the provisioning and deprovisioning of cloud resources is user-centric. Regardless of the user type, the next step is the application of a load balancing mechanism to place the offloaded task ③. For a trusted user, capacity constraint is the only condition used to admit the service request. Edge servers are tested in a round-robin manner, one at time, until a server is found to accommodate the request. If there is no edge server available to supply the computing demand, the load balancing will route it to the back-up cloud, where its admission depends on the idle capacity in this layer. For an untrusted user, the system additionally verifies the threshold condition, i.e., whether or not the number of untrusted user in the targeted computing node (edge server or back-up cloud) is below a threshold. If and only if both conditions are valid, the service request will be accepted. When the offloaded task is completed, the user releases the used VM independently of the cloud layer it is. In this respect, if the departure takes place in the MEC ④, the orchestrator will migrate an existing user from the back-up cloud to the MEC in order to thwart the SLA violation ⑤ - priority will be given to the trusted users. Since the VM migration from the back-up cloud to an edge server represents a new admission decision in an edge server, it must abide by threshold security control in the target node. Finally, if a departure is from the back-up cloud ⑥, the system updates its resource availability.

C. SYSTEM ASSUMPTION

We assume a 5G-driven MEC deployment such as the one in Fig. 1., where users under a small cell regions A_1 and A_2 benefit from a double coverage area to offload their computing tasks to either the MEC or the back-up cloud while users within the macro-cell only region A_0 can solely rely on the latter. For the sake of risk management, the users are grouped into two categories, namely the trusted category and the untrusted category. The former comprises of well-behaved and well-known clients and connected-devices and the latter comprises of unknown or misbehaving clients and connected-devices.

From an operational standpoint, Fig. 1 shows that the NFV MANO plays a key role in leveraging a risk-aware orchestration of the cloud resources since it acts perversely across the physical plane, the virtual plane, and the security service plane. Therefore, the proposed risk mitigation method should be performed by the NFV MANO.

For a risk-centric edge server MANO, the proposed mitigation method copes with the risk reduction at the MEC sites only. In this regard, the objective of the proposed risk-centric cloud orchestration is to augment the virtual capacity and enhance the security and dependability of edge servers through a cooperation with the back-up cloud. Therefore, we assume that the NFV MANO is accountable for the location-aware provisioning and placement of VMs, the VM migration between the MEC and the back-up cloud, and the execution of security controls to avoid the risk inflation. Similar to [27]–[29], we emphasize exclusively the management and operation of the computing resources and disregard the allocation of the wireless resources.

D. RISK MANAGEMENT

1) RISK IDENTIFICATION AND INVENTORY

The MNO's assets consist of the elements of NFVI physical, virtual planes, VNFs planes, NFV MANO as well as the Service Level Agreement (SLA) which specifies that a low-latency should be exercised to ensure the proper QoS and QoE. The threats that endanger the systems are:

1) Denial of Service (DoS): an untrusted user that seeks to exhaust the MEC site resources by launching a massive number of service requests to overwhelm its capacity [30], [31]. Ultimately, a DoS aims at making the edge servers unavailable for trusted users.
2) Flash network traffic: a sudden growth in MEC resource consumption due to a large-scale event that can compromise the system availability [32].
3) SLA violation: when the edge servers are full, the system starts routing the incoming service requests to the back-up cloud with the goal to keep its high availability. The site effect of this action is an increase in the perceived latency, which damages the SLA.

The vulnerabilities of a MEC site are due to its resource constrained nature that makes it susceptible to a surge in the traffic load and lack of security polices to control the number of untrusted users who can take over the entire MEC site.

2) RISK ASSESSMENT

The role of a risk assessor is to compute a risk by identifying the probability of a successful exploit to occur and its loss impact. The probability of a successful exploit will be discussed in the Sub-section V-E while the loss impact in conjunction with the vulnerability exploitation event is shown in Table 2.

TABLE 2. Vulnerability exploitation events and loss impact.

e	Event	Loss Impact	Numeric Scale (I_e)
e_1	Violation of the SLA due to a growing latency	Very low	$I_{e_1} = 1$
e_2	Admission of an untrusted user	Low	$I_{e_2} = 2$
e_3	Exhaustion of the MEC resources due to flash network traffic	Medium	$I_{e_3} = 3$
e_4	Exhaustion the MEC resources due to a DoS attack from untrusted users	Very high	$I_{e_4} = 5$

3) RISK CONTROL

The MEC sites have been designed considering that the exceeding workload is served by a subsequent MEC layer when the computing demand spikes [29], [33]. This procedure ensures a slow increase in the latency while promoting a high service availability and dependability. However, there is a need to embed the risk-awareness into the admission control decision; otherwise, a MEC site might end up being highly populated with untrusted users who can wreak havoc a physical server or an entire MEC site. To safeguard the system, security controls that limit the presence of untrusted users and keep the risk level within acceptable levels should be put forward. Thus, the cooperation between the MEC sites and the back-up cloud must be orchestrated in such a way that low-latency, high service availability, dependability, and security are holistically accomplished and agilely operated by the NFV MANO. In this work, the following security controls are used to mitigate the risks:

- Multiple thresholds: thresholds are used to limit the number of untrusted users connected to a edge server and the back-up cloud. As for the admission decision at the MEC layer, no untrusted user will be admitted into an edge server when its occupancy matches the threshold level. This control will cause the load balancing to seek another server, which can satisfy this condition, within the same MEC site in a round-robin manner. At the back-up cloud, the threshold applies to the number of untrusted in the whole layer.
- Load balancing: due to the latency sensitiveness, users within a MEC site are primarily assigned to it. However, if the edge servers are running out of capacity, the back-up cloud will supply the computing demand. To minimize the risk of SLA violation, the location-aware load balancing will bring back the trusted and untrusted users that are served by the back-up cloud to the MEC site whenever there exists room to do so. In this respect, the trusted users have the priority while the untrusted users can only be migrated if there is no trusted user running its application at the back-up cloud and the migrating VM does not violate the occupancy threshold condition.

By taking these two security controls into consideration, the proposed cloud orchestrator reduces the risk of vulnerability exploitation due to event e_1, e_2, e_3 that are shown in Table 2 and eliminates the risk of DoS attack due to untrusted users (e_4). Fig. 3 and Fig. 4 show the application of the security controls in the provisioning and deprovisioning of cloud resources in the system under analysis based on the type of event, whether it is an arrival or departure of a service, the security profile of the user, the user location, and the serving cloud layer. Remarkably, the proposed risk-aware cloud orchestrator is agile because it responds in the same timescale of arrivals and departures by leveraging the multiple thresholds and load balancing which jointly prevent the risk from inflating in an unrestrained way.

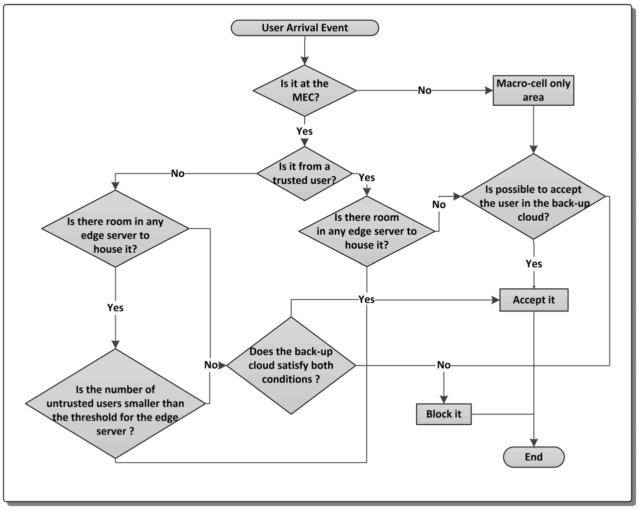
User Arrival Event
Is it at the MEC?
No
Macro-cell only area
Yes
Is it from a trusted user?
No
Yes
Is possible to accept the user in the back-up cloud?
No
Is there room in any edge server to house it?
Is there room in any edge server to house it?
Yes
Yes
Yes
Accept it
No
Does the back-up cloud satisfy both conditions ?
No
Is the number of untrusted users smaller than the threshold for the edge server ?
Block it
End

FIGURE 3. Proposed risk-aware orchestrator: user arrival decision making and controls.

E. STOCHASTIC MODEL

Let $\mathbb{C} \triangleq \{0, 1, 2, \cdots, C-1\}$ be the set of clouds where $|\mathbb{C}|$ is the cardinality of $\mathbb{C}$. The first index of $\mathbb{C}$ corresponds to the back-up cloud while the j^{th} ($1 \leq j \leq |\mathbb{C}| - 1$) ones represent the deployed MEC sites. Let N_S and S_{VM}^{E} denote the number of edge servers and VMs per edge server, respectively. The capacity of the j^{th} MEC site is given by $N_S \times S_{VM}^{E}$ VMs. Given the resourcefulness of the back-up cloud, its capacity is measured by the total number of available VMs as seen by the edge sites. In this regard, it is S_{VM}^{B} VMs.

It is assumed that the arrivals at the macro-cell only region follow a Poisson process with rate λ_b while the service time is assumed to be exponentially distributed with mean $1/\mu_b$. Recall that the proposed risk-aware orchestrator focuses on the MEC operation. In this respect, the computing demand in this region is assumed to be a background traffic. As for the j^{th} MEC site, arrivals take place according to independent Poisson processes with rates λ_j^t and λ_j^u for the trusted and untrusted user, respectively, while the service times are exponentially distributed with mean $1/\mu_t$ and $1/\mu_u$ for the

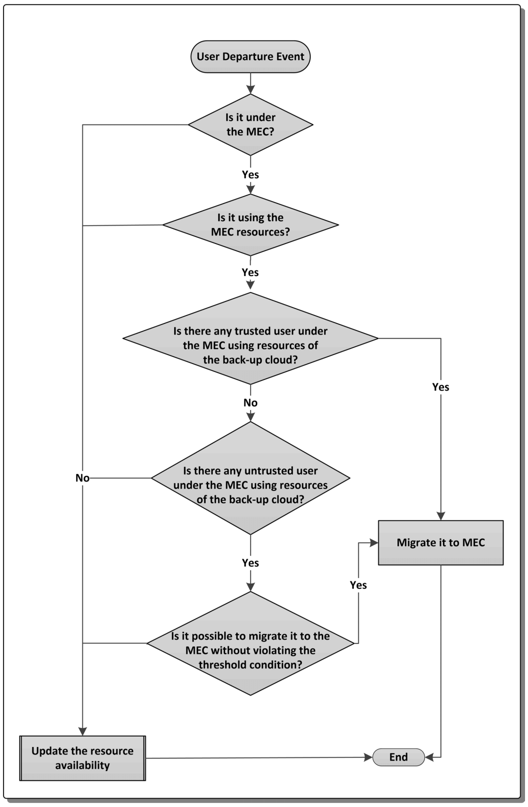
User Departure Event
Is it under the MEC?
Yes
Is it using the MEC resources?
Yes
Is there any trusted user under the MEC using resources of the back-up cloud?
No
Yes
No
Is there any untrusted user under the MEC using resources of the back-up cloud?
Migrate it to MEC
Yes
Yes
Is it possible to migrate it to the MEC without violating the threshold condition?
Update the resource availability
End

FIGURE 4. Proposed Risk-aware orchestrator: user departure decision making and controls.

respective category. The notations that are used throughout this article is summarized in Table 3.

Let the tuple $(\mathbb{S}, \mathbf{Q})$ represents the proposed multi-dimensional Markov model for the a 5G-Driven MEC deployment where $\mathbb{S}$ specifies the state space and $\mathbf{Q}$ the infinitesimal generator. Let

$$s = \Big\{s_0, s_1^t, s_1^u, \cdots, s_j^t, s_j^u, \cdots, s_{|\mathbb{C}|-1}^t, s_{|\mathbb{C}|-1}^u, s_{11}^t, s_{11}^u, \cdots, s_{ij}^t, s_{ij}^u, \cdots, s_{N_S|\mathbb{C}|-1}^t, s_{N_S|\mathbb{C}|-1}^u\Big\} \quad (1)$$

denote the state of the system in every region of the 5G network, where s_0 represents the number of VMs of the back-up cloud that are allocated to users within the macro-cell region only; s_j^t and s_j^u denote the number of trusted and untrusted VMs within the j^{th} MEC site that are using resources of the back-up cloud system. For instance, s_1^t and s_1^u represent the number of trusted and untrusted users within the first MEC site that are consuming resources of back-up cloud. Finally, s_{ij}^t and s_{ij}^u represent, respectively, the number of trusted and untrusted VMs that are using resources from the i^{th} $(1 \leq i \leq N_S)$ physical server at the j^{th} MEC. For example, s_{11}^t and s_{11}^u specify the number of trusted and untrusted users using the first edge server in the first MEC site, respectively. The scope of $\mathbb{S}$ must abide by the capacity constraints of the back-up cloud and edge servers as well as the fact that number of untrusted users in the back-up cloud and the edge server are capped by the security thresholds T and T_{ij}, respectively. Thus, $\mathbb{S}$ is given by

TABLE 3. Notations.

Symbol	Definition
$\mathbb{C}$	Set of cloud data centers.
$\lvert\mathbb{C}\rvert$	Cardinality of $\mathbb{C}$.
N_S	Number of edge servers.
S_{VM}^E	Number of VMs per edge server.
S_{VM}^B	Capacity of the back-up cloud
λ_b	Arrival rate of the background traffic.
λ_t	Arrival rate of the trusted user.
λ_u	Arrival rate of the untrusted user.
μ_b	Service rate for the background traffic.
μ_t	Service rate for the trusted user.
μ_u	Service rate for the untrusted user.
$\mathbb{S}$	State space.
$\mathbf{Q}$	Infinitesimal generator.
s	State in $\mathbb{S}$.
s_0	Number of VMs of the back-up cloud that are allocated to users within the macro-cell region only.
s_j^t	Number of trusted VMs within the j^{th} MEC site that are using resources of the back-up cloud.
s_j^u	Number of untrusted VMs within the j^{th} MEC site that are using resources of the back-up cloud.
s_{ij}^t	Number of trusted VMs that are using resources from the i^{th} $(1 \leq i \leq N_S)$ physical server at the j^{th} MEC.
s_{ij}^u	Number of untrusted VMs that are using resources from the i^{th} $(1 \leq i \leq N_S)$ physical server at the j^{th} MEC.
T	Threshold on the back-up cloud.
T_{ij}	Threshold on the i^{th} $(1 \leq i \leq N_S)$ physical server at the j^{th} MEC.
$\boldsymbol{\pi_s}$	Steady-state distribution probability.
R_T	Total security risk.

$$\mathbb{S} = \Bigg\{ s : s_0 + \sum_{j=1}^{|\mathbb{C}|-1} (s_j^t + s_j^u) \leq S_{VM}^B;\ s_{ij}^t + s_{ij}^u \leq S_{VM}^E;\ \sum_{j=1}^{|\mathbb{C}|-1} s_j^u \leq T;\ s_{ij}^u \leq T_{ij};\ 1 \leq i \leq N_S,\ 1 \leq j \leq |\mathbb{C}|-1 \Bigg\}. \quad (2)$$

Let $q(s, s')$ be an entry of the infinitesimal generator $\mathbf{Q}$ that represents the transition rate from the state s to the state s' in $\mathbb{S}$. To populate $\mathbf{Q}$ with $q(s, s')$ for all permissible states in the state space $\mathbb{S}$, the following events are considered:

1) Arrival of a background user at the back-up cloud: Upon an arrival with rate λ_b, the state variable s_0 will be incremented if the $s_0+\sum_{j=1}^{|\mathbb{C}|-1}(s_j^t+s_j^u)<S_{VM}^B$ condition holds.

2) Departure of a background user from the back-up cloud: A service completion for a background user, which happens with the service rate of $s_0\mu_b$, will decrement state variable s_0.

3) Departure of a trusted user within a MEC site using the resources of the backup cloud: A service completion for a trusted user, which happens with the service rate of $s_j^t\mu_t$, will decrement state variable s_j^t.

4) Departure of a untrusted user within a MEC site using the resources of the backup cloud: A service completion for a untrusted user, which happens with the service rate of $s_j^u\mu_u$, will decrement state variable s_j^u.

TABLE 4. Transitions from the state $s = (s_0, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u) \in \mathbb{S}$.

#	$s' \in \mathbb{S}$	$q(s, s')$	Condition
1	$(s_0 + 1, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u)$	λ_b	$s_0 + \sum_{j=1}^{\lvert\mathbb{C}\rvert-1}(s_j^t + s_j^u) < S_{VM}^B$.
2	$(s_0 - 1, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u)$	$s_0\mu_b$	$s_0 > 0$.
3	$(s_0, s_j^t - 1, s_j^u, s_{ij}^t, s_{ij}^u)$	$s_j^t\mu_t$	$s_j^t > 0$.
4	$(s_0, s_j^t, s_j^u - 1, s_{ij}^t, s_{ij}^u)$	$s_j^u\mu_u$	$s_j^t > 0$.
5	$(s_0, s_j^t, s_j^u, s_{ij}^t + 1, s_{ij}^u)$ $(s_0, s_j^t, s_j^u, s_{ij}^t, s_{lj}^t + 1, s_{ij}^u)$ $(s_0, s_j^t + 1, s_j^u, s_{ij}^t, s_{ij}^u)$	λ_t	$s_{ij}^t + s_{ij}^u < S_{VM}^E$. $s_{ij}^t + s_{ij}^u = S_{VM}^E \wedge s_{lj}^t + s_{lj}^u < S_{VM}^E$. $s_{ij}^t + s_{ij}^u = S_{VM}^E \forall i \wedge s_0 + \sum_{j=1}^{\lvert\mathbb{C}\rvert-1}(s_j^t + s_j^u) < S_{VM}^B$.
6	$(s_0, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u + 1)$ $(s_0, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u, s_{lj}^u + 1)$ $(s_0, s_j^t, s_j^u + 1, s_{ij}^t, s_{ij}^u)$	λ_u	$s_{ij}^t + s_{ij}^u < S_{VM}^E \wedge s_{ij}^u < T_{ij}$ $s_{ij}^t + s_{ij}^u = S_{VM}^E \vee s_{ij}^u = T_{ij} \wedge s_{lj}^t + s_{lj}^u < S_{VM}^E \wedge s_{lj}^u < T_{lj}$. $s_{ij}^t + s_{ij}^u = S_{VM}^E \vee s_{ij}^u = T_{ij} \forall i \wedge s_0 + \sum_{j=1}^{\lvert\mathbb{C}\rvert-1}(s_j^t + s_j^u) < S_{VM}^B \wedge \sum_{j=1}^{\lvert\mathbb{C}\rvert-1} s_j^u < T$.
7	$(s_0, s_j^t, s_j^u, s_{ij}^t - 1, s_{ij}^u)$ $(s_0, s_j^t - 1, s_j^u, s_{ij}^t, s_{ij}^u)$ $(s_0, s_j^t, s_j^u - 1, s_{ij}^t - 1, s_{ij}^u + 1)$	$s_{ij}^t\mu_t$	$s_{ij}^t > 0 \wedge s_j^t = s_j^u = 0$. $s_{ij}^t > 0 \wedge s_j^t > 0$. $s_{ij}^t > 0 \wedge s_j^t = 0 \wedge s_j^u > 0 \wedge s_{ij}^u < T_{ij}$.
8	$(s_0, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u - 1)$ $(s_0, s_j^t - 1, s_j^u, s_{ij}^t + 1, s_{ij}^u - 1)$ $(s_0, s_j^t, s_j^u - 1, s_{ij}^t, s_{ij}^u)$	$s_{ij}^u\mu_u$	$s_{ij}^u > 0 \wedge s_j^t = s_j^u = 0$. $s_{ij}^u > 0 \wedge s_j^t > 0$. $s_{ij}^u > 0 \wedge s_j^t = 0 \wedge s_j^u > 0$.

5) Arrival of a trusted user at the MEC site: The service request will be admitted in the i^{th} ($1 \leq i \leq N_S$) edge server if the condition $s_{ij}^t + s_{ij}^u < S_{VM}^E$ holds, which will increment the state variable s_{ij}^t with rate λ_t. However, if the condition fails, the l^{th} edge server will be verified. Again, the capacity constraint condition $s_{lj}^t + s_{lj}^u < S_{VM}^E$ is checked and if a true outcome is revealed, the service request will be housed at that server. If no server at the j^{th} MEC site can accept the service request, then it goes to the back-up cloud. In this case, the condition $s_0 + \sum_{j=1}^{|\mathbb{C}|-1}(s_j^t + s_j^u) < S_{VM}^B$ is tested and a positive outcome will lead to an increment of the state variable s_j^t.

6) Arrival of an untrusted user at the MEC site: The user will be hosted by the i^{th} ($1 \leq i \leq N_S$) edge server if the capacity constraint condition $s_{ij}^t + s_{ij}^u < S_{VM}^E$ and security control condition $s_{ij}^u < T_{ij}$ are true. Similar to the previous case, a round-robin scheduling is applied until an edge server that fulfills both conditions is found in the j^{th} MEC site. If no edge server satisfies the conditions, the service request goes to the back-up cloud and again the capacity constraint condition $s_0 + \sum_{j=1}^{|\mathbb{C}|-1}(s_j^t + s_j^u) < S_{VM}^B$ and the security control condition $\sum_{j=1}^{|\mathbb{C}|-1} s_j^u < T$ are verified. If the outcome is positive the state variable s_j^u will be incremented. Regardless the user placement, the transition is triggered with rate λ_u.

7) Departure of a trusted user from the MEC site: A service completion of this user will make the load balancing to search for users that are camped within the same j^{th} MEC site but who are served by the back-up cloud. If a trusted user is found, then it will be brought to an edge server, which will decrement the state variable s_j^t. However, if no trusted user is found and an untrusted user is, then it will be migrated to the MEC site if the security control condition $s_{ij}^u < T_{ij}$ is held. This event will simultaneously decrement the state variable s_j^u, increment s_{ij}^u, and decrement s_{ij}^t. If no user is found, the state variable s_{ij}^t will be decremented only. Regardless of the destination state, the transition will occur with service rate $s_{ij}^t\mu_t$.

8) Departure of an untrusted user from the MEC site: Similar to the previous event, a service completion using edge servers computing resources will make the load balancing to migrate a user from the back-up cloud to the MEC site. If a trusted user is found $s_j^t > 0$, then following changes in the state variables will occur simultaneously: s_j^t will decrement, s_{ij}^t will increment, and s_{ij}^u will decrement. If no trusted user is found and an untrusted is $s_{ij}^u > 0$, then it will be brought to the MEC. If no user is found at the back-up cloud, the state variable s_{ij}^u will be decremented with service rate $s_{ij}^u\mu_u$.

Table 4 presents the stochastic model for the proposed risk-aware orchestrator for the aforementioned events. For simplicity, it takes the state $s = (s_0, s_j^t, s_j^u, s_{ij}^t, s_{ij}^u) \in \mathbb{S}$ as the present state and formally specifies the successor state $s' \in \mathbb{S}$, the transition rate $q(s, s')$, and the conditions under which the state transition is triggered for the events previously stated. It is noteworthy that the risk-aware orchestrator follows the procedures described in Fig. 3 and Fig. 4.

Fig. 5 illustrates a small-scale state transition diagram for all transitions from and to the state $(3, 1, 0, 2, 0, 0, 1) \in \mathbb{S}$ considering a system with a MEC site with two edge servers each with capacity of 3 VMs where the threshold is 1 as well as a back-up cloud with capacity of 5 VMs with the threshold equal to 2. An arrival of the background traffic will make the system to transit to the state $(4, 1, 0, 2, 0, 0, 1)$. By the same token, a departure of such user will evolve the system to the state $(2, 1, 0, 2, 0, 0, 1)$. A departure of a trusted user within the MEC site using resources from the back-up cloud will move the system to the state $(3, 0, 0, 2, 0, 0, 1)$ while an arrival of the same category of user will be admitted into the edge server, which will cause the state transition to $(3, 1, 0, 3, 0, 0, 1)$. As for the untrusted user, its arrival in the system will be accepted into the first server that will

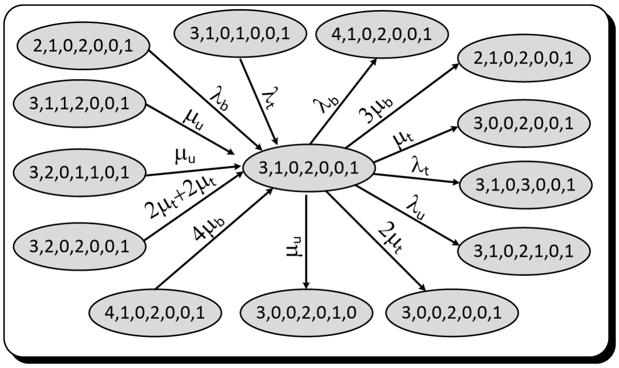

FIGURE 5. State transition diagram: $|\mathbb{C}| = 2$, $N_S = 2$ edge servers, $S_{VM}^E = 3$ VMS, $S_{VM}^B = 5$ VMs, $T_{11} = T_{21} = 1$ untrusted VMs, and $T_1 = 2$.

lead the system to (3, 1, 0, 2, 1, 0, 1). A departure of the a trusted user from the MEC site triggers the load balancing security control that will bring a trusted user from the back-up cloud to the MEC. In this respect, the final state will be (3, 0, 0, 2, 0, 0, 1). Similarly, a service completion of an untrusted user will trigger the load balancing that will cause the transition to the state (3, 0, 0, 2, 0, 1, 0). Note that there are two ways to reach the state (3, 1, 0, 2, 0, 0, 1) from the state (3, 2, 0, 2, 0, 0, 1). Despite the same service rate, the reasons are distinguishable. For first case, the trusted user using resources from the back-up cloud departs with rate $2\mu_t$ and the resource availability is updated - see Fig. 4. For the second case, a trusted user departs from the MEC site using resources from the first edge server. For this case, the load balancing opportunistically acts which decrements the number of trusted users using resources from the back-up cloud. By using the same rationale, the remaining state transitions can be similarly explained.

F. RISK METRICS

Denote $\boldsymbol{\pi_s}$ the steady-state probability of the proposed stochastic model that can be determined by solving the system of linear equations $\boldsymbol{\pi_s}\mathbf{Q} = 0$ along with the normalization condition $\sum_{s\in\mathbb{S}} \pi_s = 1$ [35]. With $\boldsymbol{\pi_s}$, the total risk can be computed as follows. The events of vulnerability exploitation e_1, e_2, and e_3, which are described in Table 2, can be formally specified considering the following subset of states:

$$e_1 = \Big\{s|s_{ij}^t + s_{ij}^u = S_{VM}^E \wedge s_j^t > 0 \vee s_j^u > 0;\, 1 \leq i \leq N_S,\\ 1 \leq j \leq |\mathbb{C}| - 1\Big\} \subseteq \mathbb{S}, \tag{3}$$

$$e_2 = \Big\{s|s_{ij}^t + s_{ij}^u = S_{VM}^E < S_{VM}^E \wedge s_{ij}^u < T_{ij};\, 1 \leq i \leq N_S,\\ 1 \leq j \leq |\mathbb{C}| - 1\Big\} \subseteq \mathbb{S}, \tag{4}$$

$$e_3 = \Big\{s|s_{ij}^t + s_{ij}^u = S_{VM}^E;\, 1 \leq i \leq N_S,\, 1 \leq j \leq |\mathbb{C}| - 1\Big\}\\ \subseteq \mathbb{S}. \tag{5}$$

Note that a state $s \in e_1$ is violating the SLA since the applications are running at the back-up cloud. By the same token, a state $s \in e_2$ is at risk of accepting an untrusted user. Finally, a state $s \in e_3$ is denying the service to users due to the MEC resource exhaustion. Let $\boldsymbol{\pi_s}$ denote the steady-state probability vector of the proposed stochastic model. With $\boldsymbol{\pi_s}$, the total security risk R_T can be computed as

$$R_T = R_{e_1} + R_{e_2} + R_{e_3}, \tag{6}$$

where $R_{e_1} = \sum_{s\in e_1} \pi_s \times I_{e_1}$, $R_{e_2} = \sum_{s\in e_2} \pi_s \times I_{e_2}$, and $R_{e_3} = \sum_{s\in e_3} \pi_s \times I_{e_3}$ are the events of vulnerability exploitation risk times their loss impact numeric scales as presented in Table 2.

VI. NUMERICAL RESULTS

We consider a 5G-driven MEC deployment with $|\mathbb{C}| = 2$, $N_S = 2$ edge servers, $S_{VM}^E = 5$ VMS, $S_{VM}^B = 20$ VMs, $T_{11} = T_{21} = 2$ untrusted VMs, $T_1 = 5$ untrusted VMs, $\mu_t = \mu_u = \mu_b = 6.6$ s^{-1}. Our analysis focuses on how the untrusted traffic adversely impacts the security risk. For this reason, we fix the background arrival rate and the trusted user arrival rate as $\lambda_b = \lambda_t = 5$ service requests/s while λ_u varies. For performance comparison, we consider as baseline approach a security risk-unaware orchestrator scheme, which ultimately represents the traditional MEC resource allocation schemes, where the exceeding workload is shifted to a more resourceful cloud layer [29], [33].

Fig. 6 shows that even though both schemes depart from the same security risk level, the proposed risk-aware orchestrator is able to mitigate the total risk and its individual components more successfully as the arrival rate of untrusted users raises while the opposite trend is observed for the risk-unaware orchestrator. To a large extent, this outcome is a consequence of the risk due to the admission of untrusted users. Despite the fact that both schemes yield a reduction in this indicator as the traffic goes up, the proposed scheme presents a steep drop against a slow one for the baseline method. Overall, the reason for this result is that under light traffic load, untrusted users have more opportunities to connect to the MEC which elevates the risk level. Conversely, from medium to heavy traffic loads, resources become more and more scarce and the rejection of untrusted users takes place more often, which deflates the risk. Since the proposed scheme still uses the threshold at the edge servers to limit the presence of untrusted users, it mitigates the risk more efficiently. Remarkably, the drop in the risk of admission of untrusted users from $\lambda_u = 8$ to 10 service requests/sec is offset by an increase in risks of SLA violation and resource exhaustion due to a flash network traffic or a DoS attack from untrusted users. Consequently, the total risk for the baseline algorithm continues to raise. However, thanks to the proposed security controls, the risk-aware orchestrator realizes a less riskier operation. An analysis of the these individual risks reveals that the use of a threshold gives more room to trusted users to enjoy the cloud resources at the edge which decreases the likelihood of servers' resource exhaustion. Additionally, it also prevents the DoS attack from untrusted users from happening, which

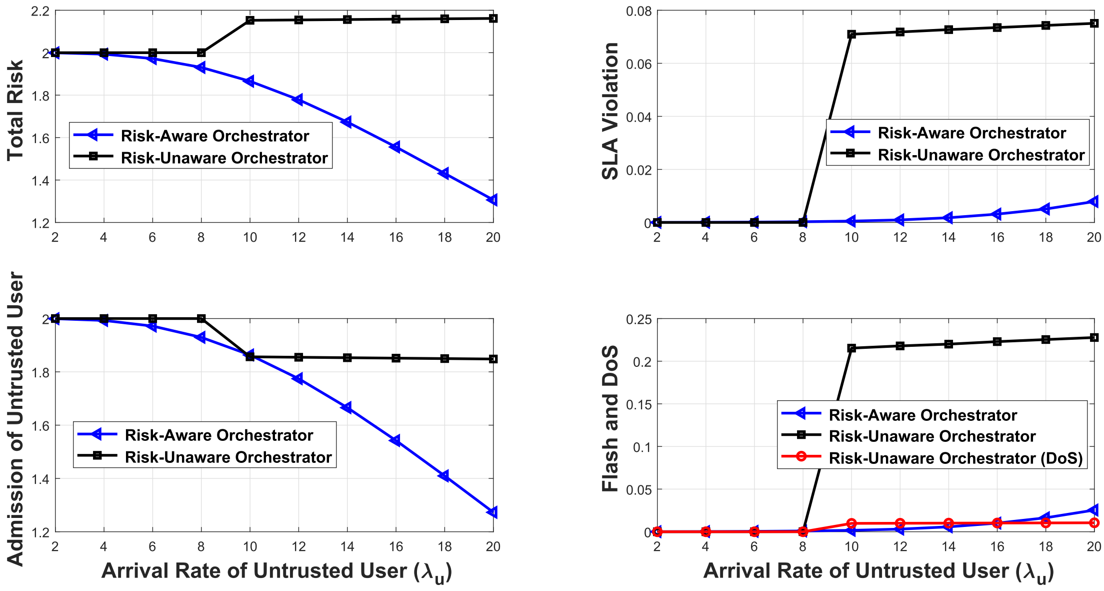

FIGURE 6. Risk profile for edge servers against an increase in the untrusted traffic load.

is in turn perceived by the baseline orchestrator. To illustrate the impact of resource exhaustion in the risk level due to the flash network traffic only, for $\lambda_u = 20$ service requests/sec, the baseline algorithm is nine times more riskier than the proposed scheme. Notably, despite their rejections at the MEC, untrusted users can use cloud services by connecting to the back-up cloud remotely. As expected, with the increase in the traffic load, more users start being served by the back-up cloud that leads to an increase in the risk of SLA violation. However, the proposed scheme benefits from the location-aware load-balancing security control to bring these users back to the MEC which ultimately drops this risk while the baseline method keeps violating the SLA during the entire service duration. To summarize, the proposed risk-aware edge server orchestrator is able to mitigate the total risk at which the MEC sites are exposed to by tackling individual risk components efficiently while eliminating the risk of a DoS attack.

VII. CONCLUSION

Agile security is a process that if taken by design will empower MNOs to deploy, manage, and operate a less risky 5G network. In this article, we have shown how agile security can leverage NFV to achieve such level of security and how it can be integrated in the daily activities of MNOs. As a process, agile security should rely on a risk management framework in order to raise the visibility of the known threats when it comes to the commitment of the cloud resources. To illustrate this concept, we have proposed a agile security risk-aware edge server orchestrator for 5G-driven MEC deployment. We have also shown that this scheme is able to keep the total risk under manageable levels by applying some security controls such as multiple thresholds and a load balancing security control.

Potential directions for future research include:

- **Security-aware edge server scheduling:** the proposed scheme assigns the untrusted user to a server in a round-robin fashion. However, a more secure approach would be to intelligently observe the distribution of untrusted users in a server and make the assignment accordingly.
- **Moving target defense (MTD):** malicious users who impersonate themselves as a untrusted users might aim at gaining control of the hypervisors by compromising the VMs. In this sense, acquiring the knowledge of the edge servers is key to launch the attack. To mitigate such risk, MTD might be applied to deceive adversaries by either changing the networking characteristics of the targeted site or switching the untrusted users among servers across all cloud layers. Proactive monitoring discussed in Section III is instrumental for the success of this defense strategy since red teams can test it to prove its effectiveness against the adversaries.
- **Optimal probabilistic risk-aware:** there is a number of uncertainties around 5G systems including players, architecture, services, and applications, to name a few. In a cyberspace, uncertainties traditionally lead to new threat agents and vulnerabilities that attract the attackers. To minimize the risk at long term, MNO might invest in optimal probabilistic risk-aware orchestrator, which makes decisions on an ongoing basis taking into account the system stochasticity. In this context, sequential decision making tools such as Markov decision process and reinforcement learning arise as potential techniques to accomplish such objective.

REFERENCES

[1] M. Pattaranantakul, R. He, Q. Song, Z. Zhang, and A. Meddahi, "NFV security survey: From use case driven threat analysis to state-of-the-art countermeasures," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3330–3368, 4th Quart., 2018.

[2] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.

[3] G. H. S. Carvalho, I. Woungang, A. Anpalagan, M. Jaseemuddin, and E. Hossain, "Intercloud and HetNet for mobile cloud computing in 5G systems: Design issues, challenges, and optimization," *IEEE Netw.*, vol. 31, no. 3, pp. 80–89, May 2017. %bibitemNFVSec:Pattaranantakul

[4] S. Lal, T. Taleb, and A. Dutta, "NFV: Security threats and best practices," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 211–217, Aug. 2017.

[5] M. Liyanage, I. Ahmad, A. B. Abro, A. Gurtov, and M. Ylianttila, *A Comprehensive Guide to 5G Security*. Hoboken, NJ, USA: Wiley, 2018.

[6] *Common Vulnerabilities and Exposures (CVE)*. Accessed: Feb. 7, 2020. [Online]. Available: https://cve.mitre.org/index.html

[7] *Anti-Phishing Working Group (APWG) Q1 2020 Phishing Activity Trends Report*. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf

[8] K. Djemame, D. Armstrong, J. Guitart, and M. Macias, "A risk assessment framework for cloud computing," *IEEE Trans. Cloud Comput.*, vol. 4, no. 3, pp. 265–278, Jul. 2016.

[9] H. Abrar, S. J. Hussain, J. Chaudhry, K. Saleem, M. A. Orgun, J. Al-Muhtadi, and C. Valli, "Risk analysis of cloud sourcing in healthcare and public health industry," *IEEE Access*, vol. 6, pp. 19140–19150, 2018.

[10] B. T. O'Hara and B. Malisow, *CCSP (ISC)2 Certified Cloud Security Professional Official Study Guide*. Hoboken, NJ, USA: Wiley, 2017.

[11] M. Baruwal Chhetri, A. R. M. Forkan, Q. B. Vo, S. Nepal, and R. Kowalczyk, "Towards risk-aware cost-optimal resource allocation for cloud applications," in *Proc. IEEE Int. Conf. Services Comput. (SCC)*, Jul. 2019.

[12] A. Almutairi, M. I. Sarfraz, and A. Ghafoor, "Risk-aware management of virtual resources in access controlled service-oriented cloud datacenters," *IEEE Trans. Cloud Comput.*, vol. 6, no. 1, pp. 168–181, Jan. 2018.

[13] Y. Han, J. Chan, T. Alpcan, and C. Leckie, "Using virtual machine allocation policies to defend against co-resident attacks in cloud computing," *IEEE Trans. Depend. Sec. Comput.*, vol. 14, no. 1, pp. 95–108, Feb. 2017.

[14] F. Miao, L. Wang, and Z. Wu, "A VM placement based approach to proactively mitigate co-resident attacks in cloud," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2018, pp. 00285–00291.

[15] X. Wang, L. Wang, F. Miao, and J. Yang, "SVMDF: A secure virtual machine deployment framework to mitigate co-resident threat in cloud," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1–7.

[16] G. H. S. Carvalho, I. Woungang, A. Anpalagan, and I. Traore, "Optimal security-aware virtual machine management for mobile edge computing over 5G networks," *IEEE Syst. J.*, early access, Jul. 20, 2020, doi: 10.1109/JSYST.2020.3005201.

[17] H. Liang, D. Huang, L. X. Cai, X. Shen, and D. Peng, "Resource allocation for security services in mobile cloud computing," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2011, pp. 191–195.

[18] Y. Liu and M. J. Lee, "Security-aware resource allocation for mobile cloud computing systems," in *Proc. 24th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Aug. 2015, pp. 1–8.

[19] H. Raei, E. Ilkhani, and M. Nikooghadam, "SeCARA: A security and cost-aware resource allocation method for mobile cloudlet systems," *Ad Hoc Netw.*, vol. 86, pp. 103–118, Apr. 2019.

[20] B. Huang, Z. Li, P. Tang, S. Wang, J. Zhao, H. Hu, W. Li, and V. Chang, "Security modeling and efficient computation offloading for service workflow in mobile edge computing," *Future Gener. Comput. Syst.*, vol. 97, pp. 755–774, Aug. 2019.

[21] B. Huang, Y. Li, Z. Li, L. Pan, S. Wang, Y. Xu, and H. Hu, "Security and cost-aware computation offloading via deep reinforcement learning in mobile edge computing," *Wireless Commun. Mobile Comput.*, vol. 2019, pp. 1–20, Dec. 2019, doi: 10.1155/2019/3816237.

[22] P. A. Apostolopoulos, E. E. Tsiropoulou, and S. Papavassiliou, "Risk-aware data offloading in multi-server multi-access edge computing environment," *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1405–1418, Jun. 2020.

[23] G. N. Iyer, V. Raman, A. Ks, and B. Veeravalli, "On the strategies for risk aware cloud and fog broker arbitrage mechanisms," in *Proc. 4th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Mar. 2020, pp. 794–799.

[24] H. Hui, C. Zhou, X. An, and F. Lin, "A new resource allocation mechanism for security of mobile edge computing system," *IEEE Access*, vol. 7, pp. 116886–116899, 2019.

[25] Y. Kanizo, O. Rottenstreich, I. Segall, and J. Yallouz, "Optimizing virtual backup allocation for middleboxes," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2759–2772, Oct. 2017.

[26] F. He, T. Sato, and E. Oki, "Optimization model for backup resource allocation in middleboxes with importance," *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1742–1755, Aug. 2019.

[27] W. Lu, X. Meng, and G. Guo, "Fast service migration method based on virtual machine technology for MEC," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4344–4354, Jun. 2019.

[28] H. Liang, L. X. Cai, D. Huang, X. Shen, and D. Peng, "An SMDP-based service model for interdomain resource allocation in mobile cloud networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2222–2232, Jun. 2012.

[29] Q. Li, L. Zhao, J. Gao, H. Liang, L. Zhao, and X. Tang, "SMDP-based coordinated virtual machine allocations in cloud-fog computing systems," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1977–1988, Jun. 2018.

[30] K. Salah, P. Calyam, and R. Boutaba, "Analytical model for elastic scaling of cloud-based firewalls," *IEEE Trans. Netw. Service Manage.*, vol. 14, no. 1, pp. 136–146, Mar. 2017.

[31] G. Carvalho, I. Woungang, and A. S. Anpalagan, "Cloud firewall under bursty and correlated data traffic: A theoretical analysis," *IEEE Trans. Cloud Comput.*, early access, Jun. 8, 2020, doi: 10.1109/TCC.2020.3000674.

[32] I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila, and A. Gurtov, "Overview of 5G security challenges and solutions," *IEEE Commun. Standards Mag.*, vol. 2, no. 1, pp. 36–43, Mar. 2018.

[33] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

[34] P. Wang, Z. Zheng, B. Di, and L. Song, "HetMEC: Latency-optimal task assignment and resource allocation for heterogeneous multi-layer mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4942–4956, Oct. 2019.

[35] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications*. Hoboken, NJ, USA: Wiley, 2006.

GLAUCIO H. S. CARVALHO is currently with the Department of Computer Science, Ryerson University, Toronto, ON, Canada, where he worked as a Postdoctoral Fellow. He is also a Cybersecurity Professor with the School of Applied Computing, Sheridan College Institute of Technology and Advanced Learning. His research interest includes security and performance analysis of cloudified and networked systems. He served as the Chair of the IEEE Toronto Section Signals and Computational Intelligence Joint Society.

ISAAC WOUNGANG (Senior Member, IEEE) received the Ph.D. degree in mathematics from the University of the South, Toulon-Var, France, in 1994. From 1999 to 2002, he worked as Software Engineer at Nortel Networks Corporation, Ottawa, ON, Canada. Since 2002, he has been with Ryerson University, Toronto, ON, Canada, where he is currently a Professor of computer science. His current research interests include radio resource management in next generation wireless networks, big data, the Internet of Things (IoT), and cloud computing. He has published eight books and over 90 refereed technical papers in scholarly international journals and proceedings of international conferences. He has served as the Chair of the Computer Chapter, IEEE Toronto Section, from 2012 to 2018. He has Guest Edited several special issues with various reputed journals, such as *Computer Communications* (Elsevier) and *Telecommunication Systems* (Springer).

ALAGAN ANPALAGAN (Senior Member, IEEE) is currently a Professor with the ELCE Department, Ryerson University, Canada. He served the department in administrative positions as the associate chair, the program director for electrical engineering, and the graduate program director. He directs a research group working on radio resource management and radio access and networking areas within the WINCORE Lab. He has coauthored four edited books and two books in wireless communication and networking areas. He is a Fellow of the Institution of Engineering and Technology (FIET) and the Engineering Institute of Canada (FEIC). He was a recipient of the IEEE Canada J. M. Ham Outstanding Engineering Educator Award, in 2018, the SGS Outstanding Contribution to Graduate Education Award, in 2017, the Deans Teaching Award, in 2011, and Faculty Scholastic, Research and Creativity Awards from Ryerson University. He was also a recipient of the IEEE M. B. Broughton Central Canada Service Award, in 2016, the Exemplary Editor Award from the IEEE ComSoc, in 2013, and a coauthor of an article that received the IEEE SPS Young Author Best Paper Award, in 2015. He is a Registered Professional Engineer in the province of Ontario, Canada. He serves as the Vice Chair of the IEEE SIG on Green and Sustainable Networking and Computing with Cognition and Cooperation.

ISSA TRAORE received the Ph.D. degree in software engineering from the Institut National Polytechnique de Toulouse (INPT)-LAAS/CNRS, Toulouse, France, in 1998. He has been with the faculty of the Department of Electrical and Computer Engineering, University of Victoria, since 1999. He is currently a Full Professor and the Coordinator of the Information Security and Object Technology (ISOT) Lab, University of Victoria. He is the Co-Founder of Plurilock Security Solutions Inc., a network security company which provides innovative authentication technologies, and is one of the pioneers in bringing behavioral biometric authentication products to the market. His research interests include biometrics technologies, computer intrusion detection, network forensics, software security, and software quality engineering. He is an Associate Editor of the IEEE Transactions on Information Forensics and Security (TIFS).