AI-based Autonomic & Scalable Security Management Architecture for Secure Network Slicing in B5G

Chafika Benzaïd* and Tarik Taleb*† and JaeSeung Song†
* University of Oulu, Oulu, Finland
† Sejong University, Seoul, South Korea
Emails: chafika.benzaid@oulu.fi, tarik.taleb@oulu.fi, jssong@sejong.ac.kr

***Abstract*—The vital importance of securing 5G and beyond networks while meeting their stringent performance requirements has promoted the recent shift towards fully automated and smart security management. In this paper, we introduce a novel autonomic and cognitive security management framework that empowers fine-grained zero-touch security management through different levels (i.e., network functions, sub-slice, and slice) and different administrative and technological domains. We showcase the compliance of the proposed framework with the ongoing standards (e.g., ZSM, 3GPP, and NFV) and demonstrate its feasibility by advocating for potential open source solutions to implement its functional blocks in a cloud-native service-based environment.**

***Index Terms*—Zero-touch Security Management, AI, Closed Loop, Network Slicing, NFV, MANO, Lifecyle Management, Beyond 5G, and 6G.**

I. Introduction

5G and beyond networks are promising to deliver ultra-low latency, energy efficiency, ultra-high throughput and reliability, and massive connectivity, which will facilitate and accelerate the society's digitalization [1]. To reap the full benefits of future mobile networks, the potential of Software Defined Networking (SDN), Network Function Virtualization (NFV), Multi-access Edge Computing (MEC) and Network Slicing is leveraged to design a fully software-defined, virtualized and highly automated infrastructure that is service- and context-aware [2]. However, such a design approach will not only bring flexibility and agility to the network but also lead to a complex and ever-evolving cyber-threat landscape. Thus, appropriate mechanisms to enforce and manage security in such a challenging environment without compromising on performance are vital. To achieve this goal, the recent research efforts and standardization initiatives are promoting the shift toward fully automated and smart security management. In this vein, the flexibility and dynamicity enabled by network softwarization technologies are coupled to the autonomic and cognitive capabilities empowered by emerging Artificial Intelligence (AI) and Machine Learning (ML) techniques. A notable proposal is the intelligent security architecture for 5G and beyond networks that has been recently introduced by the INSPIRE-5Gplus project[1]. The proposed architecture leverages a set of emerging trends and technologies, including Zero-touch network and Service Management (ZSM)[2], Security as a Service (SECaaS), Software Defined Security (SD-SEC) and AI/ML techniques, to enable a fully automated end-to-end smart network and service security management that empowers not only protection but also trustworthiness in managing 5G network infrastructures across multiple technological domains, such as the Radio Access Network (RAN), Core Network (CN), and Transport Network (TN) [3]. The INSPIRE-5Gplus architecture follows the key design principles of the ETSI ZSM reference architecture [4] by supporting the separation of security management concerns per domain, enabling AI-assisted security management closed loops, and adopting a service-based architecture whereby the provided security management services are exposed and dynamically consumed through an integration fabric as needed.

The separation of security management concerns per domain and the adoption of service-based and SD-SEC models allow to build robust and sustainable security measures that can adapt to dynamic changes in threat landscape and security requirements in future mobile networks. Nevertheless, limiting and centralizing the security management to the domain-level may fail in fulfilling the specific and challenging performance and security demands of the diverse services envisioned in 5G and beyond networks. To fill this gap, in this paper, we propose a novel autonomic and cognitive security management framework that extends the domain-level vision adopted in [3] to provide a fine-grained zero-touch security management by introducing intelligent closed-loops with different scopes and timescales, from the node level (i.e., Virtualized Network Functions – VNFs, Cloud-native Network Functions – CNFs, and Physical Network Functions – PNFs) to the end-to-end and inter-slice levels. The adopted fine-grained approach for security management empowers the effective and swift detection and mitigation of security threats close to the source, guaranteeing high level protection of the network and system assets (i.e., services, data and network infrastructure). In addition to its adherence to the design principles of ETSI ZSM [4], the proposed framework is compliant with the 3GPP-ETSI NFV framework for managing network slicing in NFV environments [5]. Another key contribution of our work is the

[1]https://www.inspire-5gplus.eu

[2]https://www.etsi.org/committee/1431-zsm

demonstration of the feasibility of the proposed framework by first recommending a cloud-native, service-based deployment architecture based on a set of open source enabling tools and then presenting a Proof of Concept (PoC) implementation of a part of the proposed architecture to enable the fully autonomous detection of anomalies within a network slice.

The remainder of this article is organized as follows. We first present the proposed autonomic and cognitive security management framework, describing its key functional blocks and showing their mapping toward the 3GPP-ETSI NFV framework. Then, we discuss the zero-touch security enabling means, with a focus on standardization initiatives as well as emerging open source technologies that can be leveraged to implement the functional blocks of the proposed framework. Following that, we present the testbed we built to show case the feasible implementation of a part the proposed framework, leveraging some of the advocated open source tools. Before concluding the paper, we highlight some open challenges that arise from the shift towards full automation.

II. Autonomic & Cognitive Security Management Framework

The envisioned complex and dynamic cyber-threat landscape, along with the challenging performance and security demands of diverse services shaping the future mobile networks, make fully distributed and autonomous management of security an imperative. To achieve this goal, we propose a novel autonomic and cognitive security management framework (See Fig. 1) that empowers hierarchical end-to-end security self-managing capabilities across multiple domains. The framework extends the domain vision adopted in [3] to provide a fine-grained zero-touch security management by introducing AI-powered closed-loops with different scopes, from the node level to the end-to-end and inter-slice levels. The adopted fine-grained approach in managing security through different levels (i.e., network functions, sub-slice, and slice) and different domains empowers effective and swift detection and mitigation of security threats close to the source, which prevents their proliferation in the network.

It is worth mentioning that the proposed framework adheres to the key design principles of ETSI ZSM [4] by supporting the separation of security management concerns and adopting a service-based architecture whereby the provided security management services are exposed and consumed by the authorized consumers through an integration fabric. The integration fabric enables services to register, discover, and invoke security management services. It also facilitates communication among the services and between the services and other management services. The historical data and knowledge generated and used by the different security management services are saved and provided through the data services within the domain or cross-domains. In what follows, we describe the key functional blocks of the proposed framework.

A. AI-Powered Security Closed Loops

Before delving into details about the key functional blocks of our envisioned framework, we first touch upon the concept

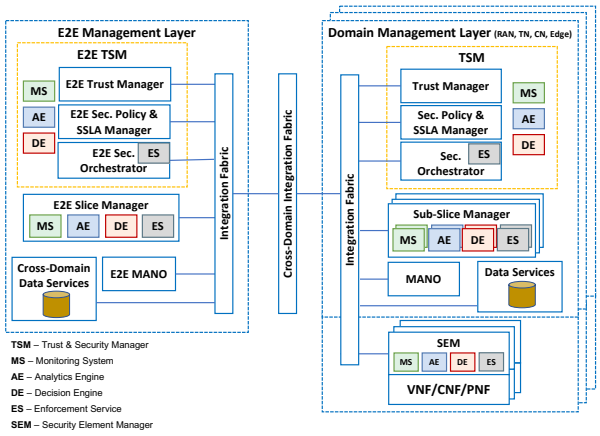

Fig. 1: The Envisioned Autonomic & Cognitive Security Management Framework.

of closed loop adopted herein. In the envisioned architecture, each closed loop is implemented using four security management functions, namely, the Monitoring System (MS), Analytics Engine (AE), Decision Engine (DE), and Enforcement Service (ES). The MS is in charge of collecting, pre-processing and reporting security-relevant data from the managed entity. The AE provides services to identify/predict potential security anomalies and attacks or determine causes of the observed security incidents based on the collected data. The DE decides the best mitigation policy needed to resolve the detected/predicted security issue to meet the desired security level. The ES translates the inferred decisions into executable actions that can be enforced on the managed entity. The ES can trigger the deployment of specific Virtual Security Functions (VSFs), such as vFirewall, vIDS, through the Management and Orchestration (MANO) platform or update the configuration of an already deployed network function (i.e., CNF, VNF, VSF, and PNF). As depicted in Fig. 1, the security management closed loops can be implemented with different scopes, ranging from the domain level to the network function level. Each network function is associated to a Security Element Manager (SEM) which is responsible for managing the security within the network function scope.

The cognition capabilities are incorporated in the closed loops by leveraging AI/ML techniques for security analytics and decision making. The cognitive level of the closed loop can be further increased by integrating the AI/ML techniques into the MS and ES to intelligently determine the relevant data to collect and decide on the actions to execute, respectively. This allows for the achievement of ultimate goal of empowering a full autonomous security management. The emerging distributed AI (DAI) techniques, including multi-agent reinforcement learning and federated learning, can be leveraged to accelerate the learning process of the AI/ML models used by the different security closed loop. The use of DAI is also expected to help in fostering data privacy preservation, as the information exchanged between the cooperating models is only limited to the model parameters without the need for exchanging any raw data [1].

While the deployed security closed loops are responsible of the autonomous handling of security within their scope, they

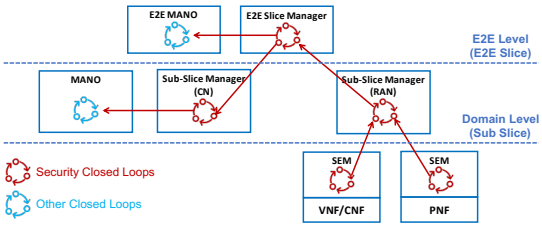

Fig. 2: An Illustrative Example of Coordination between Closed Loops.

can coordinate with other security or non-security management closed loops. The coordination between the closed loops can be performed hierarchically and/or peer-to-peer, within the same or cross-management domains. For instance, as illustrated in Fig. 2, the monitoring data and/or analytics insights from the closed loops associated to the network functions composing a RAN sub-slice can be leveraged by its associated security closed loop to identify sub-slice level security issues (e.g., detect the symptoms of a signaling DDoS attack [6]). This event may need to be reported to the E2E slice security closed loop, which will subsequently delegate the CN sub-slice security closed loop to proactively mitigate the reported issue. A decision to scale out the network functions involved in the CN sub-slice (e.g., Access and Mobility Management Function – AMF and Session Management Function – SMF) may be taken. To perform the scaling action, a coordination with, for example, the self-optimization closed loop of the domain's MANO may be required for ensuring optimal resource allocation.

B. Trust & Security Manager

As depicted in Fig. 1, the different closed loops are managed and orchestrated by the "Trust & Security Manager" (TSM) and the E2E TSM at the domain and E2E levels, respectively. The TSM, including the E2E TSM, encompasses three functional modules, namely the "Security Orchestrator", "Security Policy and SSLA Manager," and "Trust Manager". The *Security Policy & SSLA Manager* manages the SSLAs (Security Service Level Agreements) and the security policies defined by an external entity (e.g., network operator, Over-The-Top – OTT – service provider) or dynamically issued by the DEs based on the changing service and network conditions. It provides services to specify the security requirements or policies in a machine readable and structured format using a Domain Specific Language (DSL), deploys the security policies to the ES, detects and mitigates conflicts between security policies before their enforcement, and monitors the state and status of security policies as well as the fulfillment of the established SSLAs. In addition to the orchestration of the available security appliances (e.g., Firewall, DPI, IDS, and VPN) to enforce the security policies, the *Security Orchestrator* has the responsibility of designing, instantiating and managing the run-time lifecycle (e.g., activation, deactivation, and update) of the security closed loops. The instantiated closed loops or part of their components (i.e., MS, AE, DE and ES) can be dedicated to a specific managed entity or can be common to different managed entities. In the latter case, the Security Orchestrator ensures that the reusability is performed in compliance with the isolation level of slices. The *Trust Manager* continuously assesses the trustworthiness of network services and associated closed loops, their composed functions, and the hosting infrastructure. The trust score is calculated based on the trust attributes specified in the Trust Level Agreement (TLA), which may include the security measures in place, compliance with regulations (e.g., privacy preservation, operating location rules), and fulfillment of the agreed service and security levels [1].

Given the vital role played by TSMs, including E2E TSM, their compromise can be detrimental to the security and functioning of the entire network. Thus, appropriate measures need to be set up to establish trustworthy and secure interactions between the security management services within a TSM or among different TSMs, as well as with other management entities. The trust relationships should be defined and maintained in an adaptive way, taking into account the changing security context of interacting entities (e.g., security threats and risks, security policies and regulations, and applied countermeasures). Secure communication, isolation, and access control mechanisms, including identification, authentication, authorization and auditing, need to be applied and updated according to the defined trust relationship to safeguard the security management services from unavailability, misuse and unauthorized access, and prevent information leakage and damage. For more details on the security countermeasures to adopt in order to empower trust between security management services, we refer the interested readers to the authors' work in [1], [7] and the recent reports ETSI GR ZSM 010 and 3GPP TR 28.817.

C. Mapping to 3GPP-ETSI NFV Framework

In this section, we discuss how the proposed autonomic & cognitive security management framework maps towards existing standards. In particular, we consider the 3GPP-ETSI NFV framework proposed for managing network slicing in an NFV environment [5]. As illustrated in the left-hand side of Fig. 3, and according to the 3GPP terminology [8], a network slice instance (NSI), used by a network service, contains one or more network slice subnet instances (NSSIs), each of which is, in turn, composed of one or more network functions that can be managed as VNFs, CNFs and/or PNFs. The management (including lifecycle) of NSIs and NSSIs is under the responsibility of the Network Slice Management Function (NSMF) and the Network Slice Subnet Management Function (NSSMF), respectively. The network functions (i.e., VNFs, CNFs and PNFs) are managed and orchestrated using the Element Management (EM) and the NFV MANO functional blocks. The EM performs the Fault, Configuration, Accounting, Performance and Security (FCAPS) management of the network functions, while NFV MANO carries out the management of the virtualized infrastructure as well as the orchestration of resources required by the network services, VNFs and CNFs. The NFV MANO includes (i) a Virtualized Infrastructure Manager (VIM) to manage the NFVI virtual resources; (ii) a VNF Manager (VNFM), which is responsible for the NFV lifecycle management; (iii) a Container

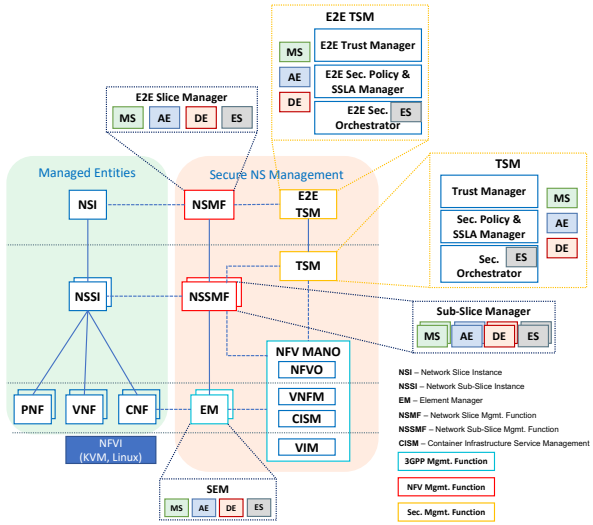

Fig. 3: Mapping of the Framework Components to the 3GPP-ETSI NFV Network Slicing Management Architecture.

Infrastructure Service Management (CISM), which is in charge of the management of containerized workloads in terms of deployment, monitoring, and lifecycle management; (iv) and a NFV Orchestrator (NFVO), which handles network resources and services by interacting with VIM and NFVM [9]. The right-hand side of Fig. 3 shows the interaction between the 3GPP slicing related management functions (i.e., NSMF and NSSMF) and the NFV architecture functional blocks (i.e., EM and NFV MANO) while illustrating how the security management functions of our framework interface as well as expand their functionalities with closed-loop security management capabilities.

III. Security Automation Means

A. Standardization Initiatives

This section provides an overview of the relevant efforts and initiatives of Standards Developing Organisations (SDOs) to foster autonomous and automated network and service management.

ETSI Generic Autonomic Network Architecture (ETSI GANA) is an architectural reference model for autonomic networking, cognitive networking and self-management [10]. GANA combines the main concepts from well-known closed-loop models of autonomic networked systems, such as Monitor-Analyze-Plan-Execute (MAPE) and FOCALE. The GANA reference architecture uses a Knowledge Plane to autonomously support the various management and control systems, including the NFV Orchestrator, SDN controller, and end-to-end service orchestrator. Recently, a proof-of-concept for applying the GANA reference model to enable end-to-end autonomic security management and control for 5G slices was proposed [11].

ETSI Experiential Network Intelligence (ETSI ENI[3]) defines a Cognitive Network Management architecture using closed-loop AI mechanisms based on context-aware policies to enable timely and actionable decisions. The architecture adopts the "Observe-Orient-Decide-Act" (OODA) closed loop model.

ETSI ZSM[4] has specified a reference architecture [4] for supporting fully-automated, end-to-end management of emerging and future networks and services. Unlike ETSI ENI which focuses on AI techniques, policy management and closed-loop mechanisms, ETSI ZSM aims at automation techniques, full automation and service management functions. ETSI ZSM is currently focusing on the specification of enablers for closed-loop management and coordination.

3GPP introduced NWDAF (Network Data Analytics Function) [12] and MDAS (Management Data Analytics Service) [13] to support network data analytics at the control and management plane, respectively. NWDAF is part of the 5G Core Network architecture. The functionalities provided by NWDAF include: (i) data collection from network functions (NFs), application functions (AFs), and OAM (Operations, Administration and Maintenance); (ii) analytics information provisioning to NFs and AFs; and (iii) ML model training and provisioning. MDAS provides data analytics related to NF, NSSI, and/or NSI. 3GPP has also specified the use cases, requirements and management services for closed loop communication service assurance in RAN and CN.

B. Zero-touch Security Enabling Technologies

The overall objective of this section is to demonstrate the feasibility of our envisioned framework, leveraging a set of open source solutions to empower zero-touch security management in 5G and beyond networks. The explored solutions include tools to: (i) enable cloud-native, service-based security management, (ii) automate security closed-loops governance, and (iii) build the monitoring and analytics services of the MS and AE security management functions.

1) Cloud-Native & PaaS Platforms: A cloud-native architecture, which is generally based on stateless micro-services deployed as containers, is recognized as the best suited technology to deliver the requisite cost-efficiency, flexibility and scalability in operating and managing 5G and beyond networks and services [14]. Cloud native design is an approach that takes full advantage of the cloud computing model to enable faster service launch and network management automation. Platform-as-a-Service (PaaS) is a key layer in a cloud-native architecture which provides a platform that allows developers to implement, run, and manage different applications without dealing with the complexity of setting up and maintaining the cloud infrastructure. Thus, offering network functions and (security) management services as CNFs that can be instantiated inside a PaaS and can expose capabilities through common and open APIs allows to their rapid deployment, upgrade, and scaling to cater to the stringent performance and security demands of emerging and future services [9]. Kubernetes[5] is becoming a de-facto standard for the deployment and orchestration of containerized applications, thanks to its built-in scalability, high availability, and fault tolerance features. In mobile networks, some of the Edge and Open RAN use cases

[3]https://www.etsi.org/committee/eni

[4]https://www.etsi.org/committee/zsm

[5]https://kubernetes.io

are already adopting Kubernetes as a platform to deploy and operate CNFs.

2) Integration Fabric: As explained in the previous section, the integration fabric facilitates the inter-operation and communication between management services, within and across domains, by providing functionalities to register, expose capabilities, discover, and invoke security management services by authorized consumers. According to [4], the Integration Fabric should support both synchronous and asynchronous communications using the request-response and publish/subscribe communication models, respectively. A combination between a service mesh solution such as Istio[6] or linkerd[7] and an event streaming platform, such as Apache Kafka[8], allows to implement the Integration Fabric functionalities. A mesh service manages inter-service traffic for synchronous communications while bolstering security and enabling observability. Meanwhile, an event streaming platform handles asynchronous communications between applications and services through event brokers. It provides the capabilities to ingest, store, process and react to a massive influx of real-time streams of data in a scalable and resilient manner. The use of an event streaming platform is important for security use cases, including real-time monitoring, analytics and reaction/prediction of security threats on the fly.

Istio is the micro-service mesh solution that has the most features and flexibility than any existing open source service mesh solutions by far. Istio supports heterogeneous environments, including Kubernetes and Virtual Machines (VMs), and multi-domains setting. Furthermore, it enables intelligent routing and load balancing between services and provides tracing, monitoring and logging features to get insights into the service mesh deployment. Finally, Istio is the undisputed leader when it comes to security features, providing a comprehensive security solution that encompasses throttling, strong identity, powerful policy, transparent TLS encryption, and authentication, authorization and audit (AAA) tools to protect both services and data exchanged between them.

Kafka is the most popular open source distributed event streaming platform, owing to its excellent performance, elasticity, low latency, fault tolerance, and high throughput. Kafka uses topics to which producers publish data and consumers subscribe to access data. Kafka can be deployed on bare-metal hardware, VMs or containers.

As illustrated in Fig. 4, we consider a combination of the capabilities of Istio and Kafka as a potential candidate to implement the Integration Fabric for a cloud-native service-based architecture.

3) Management, Orchestration & Closed Loop Automation: Open Network Automation Platform (ONAP[9]) and Open Source MANO (OSM[10]) are the most popular open source management and orchestration platforms. ONAP provides a unified framework for real-time, policy-driven orchestration, management and automation of network and edge computing services. The ONAP ecosystem includes different sub-systems that support closed loop automation, namely: (i) POLICY, which provides the capability to create and validate policies as well as identify and resolve conflicts between policies; (ii) CLAMP, which is a platform for designing and managing closed control loops; and (iii) DCAE (Data Collection, Analytics and Events), which is a platform for data collection and analysis. OSM is an ETSI NFV compliant MANO capable of modeling and automating the full life-cycle of network functions (i.e., CNFs, VNFs, and PNFs), network services and network slices. OSM's modules that enable closed-loop automation include: (i) MON, which is a monitoring module that leverages existing monitoring tools to collect metrics from VNFs and underlying infrastructure; and (ii) POL, which is a policy management module designed around the auto-scaling use case.

Another simple, yet powerful, open-source tool that comes into play to enable production-grade automation in a cloud-native environment is Ansible[11]. It is a popular automation configuration engine that uses playbooks to handle not only Day-0 provisioning tasks, but also Day-1 and Day-2 configurations of both the infrastructure and services running above it. Ansible's agentless nature and human readable language (i.e., Ansible playbooks are written in YAML) makes it an ideal tool to augment NFVM and NFVO's automation capabilities in managing and orchestrating network functions and services. It is worth mentioning that Ansible is supported by both ONAP and OSM. Moreover, Ansible is backed up and embraced by major Telco network vendors, such as Ericsson, Huawei and Nokia for automating the provisioning and configuration of their cloud-native 5G infrastructure.

4) Monitoring as a Service: Prometheus[12] and ELK Stack[13] are common open source monitoring tools used for collecting application specific metrics and logs from distributed systems. Prometheus is a time-series event monitoring tool for cloud-native, containerized environments. It uses a pull approach for gathering infrastructure- and service-level performance metrics that are collected through exporters. The Thanos[14] tool enables the use of Prometheus at large scale through multiple domains with long-term storage capabilities. Even though ELK Stack allows metrics collection, it is mainly specialized in collecting, aggregating and processing logs. ELK Stack follows a push-based model for data collection. By adopting Prometheus and ELK Stack together, it is possible to build an efficient and scalable monitoring system that can be delivered as a service for cloud-native environments.

5) Analytics as a Service: In this section, we investigate the capabilities of two open source platforms for providing AI and analytics services, namely Platform for Network Data Analytics (PNDA)[15] and Acumos AI Platform[16].

PNDA is an open source, scalable big data analytics platform for networks and services that brings together a number

[6]https://istio.io
[7]https://linkerd.io
[8]https://kafka.apache.org
[9]https://www.onap.org
[10]https://osm.etsi.org
[11]https://www.ansible.com
[12]prometheus.io
[13]www.elastic.co
[14]https://thanos.io
[15]http://pnda.io
[16]https://www.acumos.org

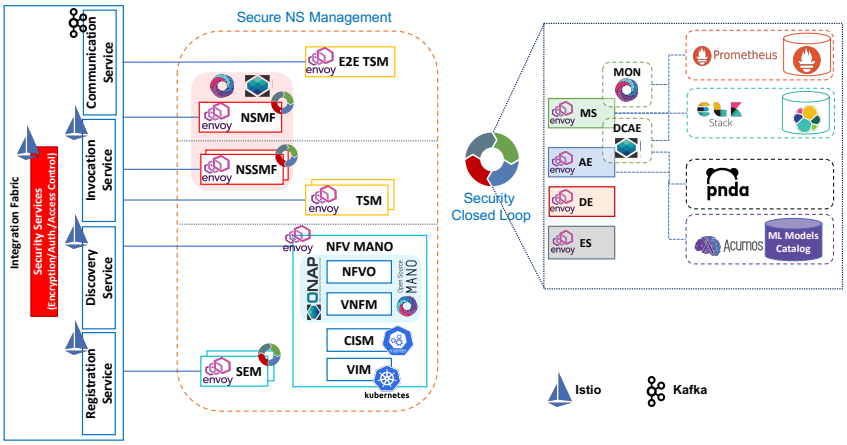

Fig. 4: A Cloud-Native Autonomic Security Management Framework Deployment.

of open source technologies (e.g., Kafka, Hadoop, and Spark). It allows to build applications for predictive analysis on time-series and deep learning applications using high-dimensional data. PNDA has been used to enable closed loop control for an ETSI NFV environment. Furthermore, ONAP is integrating PNDA as part of DCAE to provide its analytics services to the ecosystem. It is worth mentioning that ETSI ENI is currently exploring opportunities to reuse and re-purpose some PNDA tools to support more generic analytics and AI mechanisms in ENI Release 2 specifications.

Acumos AI Platform[17] is an open source framework to build, share and deploy AI/ML models. It supports multiple ML learning libraries (e.g., scikit-learn and TensorFlow). It is capable of packaging ML models into portable containerized microservices. An "Acumos-DCAE Adapter" is developed to integrate ML models from an Acumos catalogue to ONAP DCAE. ETSI ENI is considering the Acumos approach to knowledge engineering and AI algorithms in the development of Release 2 of the ENI system architecture.

C. A Cloud-Native Autonomic Security Management Framework Deployment

In this section, we provide a potential cloud-native deployment of the proposed autonomic & cognitive security management framework using the open source enabling tools introduced in the above sections. Fig. 4 illustrates the suggested deployment architecture. We consider a Kubernetes based cloud-native environment where Kubernetes is running in the cloud or on a bare-metal (i.e., no virtualization layer is needed). The management functions as well as the managed network functions are deployed as loosely-coupled container-based or VM-based services that run on Kubernetes. Running both VM-based and container-based services within a Kubernetes cluster allows to build a unified orchestration platform that will not only facilitate orchestration needs but will also foster the smooth move from VNFs to CNFs. It is worth mentioning that open source tools, such as KubeVirt[18] and Virtlet[19], allow for running VM workloads in a cloud native environment. In such a setup, Kubernetes can act as both VIM and CISM. The functions of NSMF, NSSMF, NFVO and NFVM can be provided by either ONAP or OSM. The MS and AE functions can be implemented using the MON and DCAE modules from OSM and ONAP, respectively, or by directly using the open source monitoring tools (e.g., Prometheus and ELK) and analytics platforms (e.g., PNDA and Acumos).

The management functions deployed as services interact and collaborate through the integration fabric, which is implemented by combining the capabilities of Istio and Kafka. The management functions are connected to each other using Envoy sidecar proxies to form a service mesh managed by Istio. The synchronous communication between services can be enabled through Istio, while the asynchronous communication can be performed via Kafka.

IV. Proof of Concept - Service-based Autonomic Anomaly Detection System

This section presents the testbed we built to demonstrate the feasibility of a part of our proposed architecture, namely a service-based autonomic security management system. Indeed, given the importance of the monitoring and analytics capabilities, as envisioned in the proposed architecture, in effectively detecting and mitigating security threats, the testbed aims at providing a PoC implementation of the monitoring (MS) and analytics (AE) functions as services to allow distributed and fully autonomous detection of anomalies within a network slice.

As illustrated in Fig. 5, the testbed consists of two OpenStack cloud platforms interconnected using a secure communication channel. A Kubernetes (K8s) cluster with one master node and three worker nodes is set up using four VMs managed by OpenStack. As a case study and similar in spirit to the work in [15], we consider a virtual Content Delivery Network (vCDN) service deployed as a slice at the edge. In our implementation, the vCDN slice is composed of two CNFs, namely a video streamer and a cache, chained together to provide an HTTP-based live network streaming service. The two CNFs are deployed as K8s services running a NGINX web server, and are distributed along two worker nodes. Note that only the streamer service is exposed to the end user for

[17]https://www.acumos.org
[18]https://kubevirt.io
[19]https://docs.virtlet.cloud

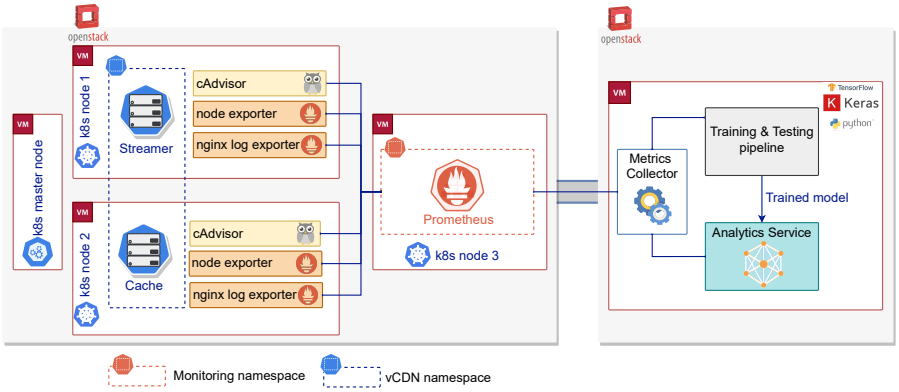

Fig. 5: A Service-based Autonomic Anomaly Detection System for vCDN Slices.

content delivery. To ensure resource isolation between slices, each vCDN slice instance has its own namespace. A fifth VM, running on the second OpenStack cloud, is used to deploy the monitoring and analytics services for anomaly detection. In this experiment, we focused on detecting application-layer DDoS attacks against a vCDN slice by identifying anomalies in the resource usage and performance metrics of vCDN's CNFs and their hosting nodes. The VM also serves as a platform for training and testing the ML models to integrate in the analytics service. To this end, the open-source tools Keras, TensorFlow and Python have been installed on the VM to create the training and testing pipeline. Both monitoring and analytics functions are deployed as containers and expose their services via RESTful APIs.

The monitoring service includes a "Metrics Collector" implemented using Python, which leverages Prometheus API to extract metrics relevant to anomaly detection. For this purpose, Prometheus relies on NGINX-to-Prometheus log file exporter, cadvisor [20], and node-exporter to scrape metrics related to NGINX server, vCDN slice's CNFs and their hosting Worker nodes, respectively. The Metrics Collector offers the capabilities to generate on-demand the dataset for training and testing the anomaly detection model or continuously collect the metrics values observed in the last x time steps in order to be fed into the Analytics Service for real-time anomaly detection. Note that the metrics are extracted as time series in CSV files.

The analytics service integrates an anomaly detection model built using the unsupervised deep learning technique LSTM (long short-term memory) AutoEncoder on multivariate time series. The use of multivariate time series allows to capture the correlation between different metrics, which results in enhanced anomaly detection accuracy. The LSTM-based AutoEncoder model is trained to reconstruct time-series for normal behavior. The inputs to the model are resource usage (e.g., CPU usage, system load, memory usage, I/O network traffic) and performance (e.g., HTTP response time) metrics. The model is trained on the time series for normal behavior of vCDN. To this end, we used a dataset of 2361 samples, where 20% of samples are held out for validation. The model is trained using 30 epochs, a batch size of 50, and Mean Absolute Error (MAE) as loss function. An anomaly is detected if the reconstruction error is above a given threshold. In this work, we used a static threshold defined as 99% of the loss distribution. The results depicted in Fig. 6 show the effectiveness of the proposed model in autonomously detecting anomalies (red dots) related to application-layer DDoS attacks against the video streamer CNF. The attacks are launched using Hulk and Slowloris tools at specific time periods, represented by the highlighted red regions in Fig. 6.

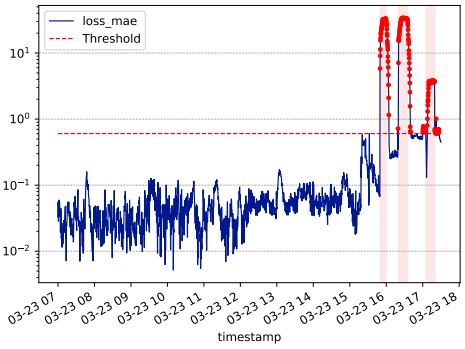

Fig. 6: Anomaly detection for vCDN's video streamer. The two first highlighted red regions correspond to Hulk attacks, while the last region represents the Slowloris attack.

It is worth noting that the testbed is a work in progress. We are currently integrating our testbed with AI4EU[21]; a platform built upon Acumos, to build and deploy AI models to incorporate in the analytics service. Furthermore, we are developing the DE and ES functions to autonomously prevent malicious CNF auto-scaling requests caused by application-layer DDoS attacks based on anomalies detected in resource usage and performance metrics.

V. Open Challenges

Despite its benefits, the shift towards full automation does not come without risks and challenges. Indeed, new attack vectors can be introduced by the different technologies and concepts leveraged to enable full automation, including virtualization, programmability, closed-loop, and AI/ML [7], [6].

[20]https://github.com/google/cadvisor

[21]https://aiexp.ai4europe.eu/#/home

Thus, appropriate measures to address security issues brought by those enablers is paramount to foster confidence in network automation.

The adoption of the PaaS paradigm to deliver (security) management services introduces both security and trust challenges. As aforementioned, the sharing of the security closed loops or part of their components may lead to violation of slice isolation. Furthermore, it is important to ensure the trustworthiness of the PaaS provider and the services running on top of the PaaS. Blockchain is a promising technology to promote trust in PaaS.

Further research efforts are also required to devise mechanisms for empowering trustworthy collaboration between closed-loops. In fact, the interaction and exchange of information among closed loops call for solutions to guarantee the accuracy and integrity of the shared information. Moreover, approaches to avert potential privacy leakage from the exchanged information are crucial, particularly when the interacting closed loops are under the control of different administrative domains. Federated Learning (FL) is a promising candidate to tackle the privacy issue, thanks to its ability to allow knowledge sharing among interacting entities without exchanging raw data. Nevertheless, how to apply FL paradigm to strategically enable hierarchical and/or peer-to-peer collaboration between closed loops, taking into account the dynamicity in deployed closed loops, the resource constraints and the stringent performance requirements, is a key challenging issue that needs careful investigation. The security of FL is another hurdle to overcome before its benefits can really be reaped. Indeed, FL is vulnerable to privacy leakage, where an adversary, including an honest-but-curious entity involved in the FL process, can carry out membership inference attacks against other entities to infer their private local data leveraging the shared model parameters. Furthermore, FL is prone to poisoning attacks, where an adversary may upload false or low-quality local model updates to impair the accuracy of the global model, which may put into peril both network's performance and security. Blockchain and Trusted Execution Environments (TEEs) are two emerging technologies that can be leveraged to strengthen FL security. The inherent decentralization and immutability properties of Blockchain technology makes it a promising solution to defeat poisoning attacks against FL models. Besides ensuring the integrity of the local and global model updates, blockchain's smart contracts can be used to identify malicious entities by automatically evaluating the quality of their local model updates against a validation dataset. Nevertheless, how to create and update the validation dataset is still an open question. The integrity and confidentiality features endowed with applications run and data saved inside TEEs make those environments a potential enabler for privacy-preserving FL. In fact, the local and global models' codes and updates as well as the aggregation algorithm used to compute the global model updates can be saved and operated over encrypted data inside the TEE. However, realizing TEE-empowered FL to protect against privacy attacks, while considering the limited memory of TEEs and the additional computation overhead engendered by encryption/decryption operations is still an open challenge.

VI. Conclusion

This paper presented a novel autonomic and cognitive security management framework that empowers fine-grained zero-touch security management for network slicing in future mobile networks. We showed its compliance with ongoing standards initiatives (i.e., ZSM, 3GPP, and NFV) and advocated for potential open source solutions that can be leveraged to implement its functional blocks in a cloud-native service-based environment, hence proved the feasibility of the envisioned framework. Despite the growing interest, the journey towards AI-powered full automation of network and service management has just started, and several challenges are still to be addressed as pointed out in this paper. It is all the hope of the authors that the elements highlighted in this paper would stimulate and shape up further research efforts, among the academic and industrial communities, to cope with these challenges.

Acknowledgment

This work was supported in part by the Academy of Finland Project 6Genesis Flagship (Grant No. 346208), and the European Union's Horizon 2020 research and innovation programme under the INSPIRE-5Gplus project (Grant No. 871808) and the CHARITY project (Grant No. 101016509). Prof. Song was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Grant No. 2020-0-00959, IoT Platform for Boosting On-device AI in 5G Environment).

References

[1] C. Benzaid, T. Taleb, and M. Z. Farooqi, "Trust in 5G and Beyond Networks," *IEEE Network Magazine (Early Access), doi: 10.1109/MNET.011.2000508*, 2021.

[2] C. Benzaid and T. Taleb, "AI-driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions," *IEEE Network Magazine*, vol. 34, no. 2, pp. 186 – 194, March/April 2020.

[3] C. Benzaid et al., "White Paper: Intelligent Security Architecture for 5G and Beyond Networks," *INSPIRE-5Gplus*, Nov. 2020.

[4] ETSI GS ZSM 002, "Zero-touch Network and Service Management (ZSM); Reference Architecture," Aug. 2019.

[5] ETSI GR NFV-EVE 012 V3.1.1, "Report on Network Slicing Support with ETSI NFV Architecture Framework," Dec. 2017.

[6] C. Benzaid and T. Taleb, "AI for Beyond 5G Networks: A Cyber-SecurityDefense or Offense Enabler?" *IEEE Network Magazine*, To appear.

[7] ——, "ZSM Security: Threat Surface and Best Practices," *IEEE Network Magazine*, vol. 34, no. 3, pp. 124 – 133, May/June 2020.

[8] 3GPP TR 28.801 V15.1.0, "Study on Management and Orchestration of Network Slicing for Network Generation Network (Release 15)," Jan. 2018.

[9] ETSI GR NFV-IFA 029 V3.3.1, "Report on the Enhancements of the NFV Architecture towards "Cloud-native" and "PaaS"," Nov. 2019.

[10] ETSI TS 103 195-2 V1.1.1, "Generic Autonomic Network Architecture; Part2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management," May 2018.

[11] TSI GANA, "White Paper No. 6: Generic Framework for Multi-Domain Federated ETSI GANA Knowledge Planes (KPs) for End-to-End Autonomic (Closed-Loop) Security Management & Control for 5G Slices," , June 2020.

[12] 3GPP TS 23.288 V17.0.0, "Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 17)," , March 2021.

[13] 3GPP TS 28.533 V16.7.0, "Management and Orchestration; Architecture Framework (Release 16)," , March 2021.

[14] 5G-PPP Software Network Working Group, “Cloud-Native and Verticals’ Services; 5G-PPP Projects Analysis,” Aug. 2019.
[15] T. Taleb, P. Frangoudis, I. Benkacem, and A. Ksentini, “CDN Slicing over a Multi-Domain Edge Cloud,” *IEEE/ACM Trans. Mobile Computing*, vol. 19, no. 9, pp. 2010 – 2027, Sep. 2020.