An evolved security architecture for distributed Industrial Automation and Control Systems

L. Rosa[1], J. Proença[1], J. Henriques[1,2], V. Graveto[1], T. Cruz[1], P. Simões[1], F. Caldeira[1,2], E. Monteiro[1]
[1] Department of Informatics Engineering, University of Coimbra, Portugal
{lmrosa, jdgomes, jpmh, vgraveto, tjcruz, psimoes, fmanuel, edmundo}@dei.uc.pt
[2] Polytechnic Institute of Viseu, Portugal

Abstract: Over the recent years, control and sensor systems used for IACS (Industrial Automation and Control Systems) have become more complex, due to the increasing number of interconnected distributed devices, sensors and actuators. Such components are often widely dispersed in the field – this is the case for micro-generation (wire-to-water generation, solar or wind), smart metering, oil and gas distribution or smart water management, among others. This IoT (Internet of Things)-centric IACS paradigm expands the infrastructure boundaries well beyond the single or aggregated-plant, mono-operator vision (mostly associated with geographically constrained systems topologies), being dispersed over a large geographic area, with increasingly small areas of coverage as we progress towards its periphery.

This situation calls for a different approach to cyber threat detection, which is one of the most relevant contributions of the ATENA (*Advanced Tools to assEss and mitigate the criticality of ICT components and their dependencies over critical infrAstructures*) H2020 project (ATENA 2016). This paper presents and describes the ATENA cyber-security architecture, designed for the emerging generation of distributed IoT IACS, leveraging technologies such as Software Defined Networking/Network Function Virtualization and Big data event processing) within the scope of a cyber-detection architecture designed to deal with the inherent challenges of dispersed IACS, involved different operator domains.

Keywords: Critical Infrastructure Protection, Industrial Automation and Control Systems, Big Data, Forensics

1. Introduction

Over the recent years, systems adopted in Critical Infrastructures (CI), such as smart grids, water, oil and gas distribution networks, have been becoming more complex due to the increasing number of interconnected distributed devices, sensors and actuators, often widely dispersed in the field, and the larger amount of information exchanged among system components. Such systems need to be flexibly and securely managed, monitored and configured, while preventing risks both from operational errors and from cyber-attacks, intrusions and malware, compromising their operation or even resulting in disasters (Edwards 2014).

Moreover, with the emergence of the IoT generation of IACS, the boundaries of the protected infrastructures expand well beyond the single or aggregated-plant, mono-operator vision. Instead of a monolithic system, deployed on geographically constrained spaces, these systems are characterized by a considerable degree of capillarity, being dispersed over a large geographic area, with increasingly small areas of coverage as we progress towards its periphery. This poses a challenge because, as the boundaries of the IACS expand towards households, they involve several other operators, such as telecommunications or utility providers, in a scenario that naturally demands the introduction of multi-tenancy mechanisms for supporting Machine-to-Machine (M2M) communications and infrastructure orchestration.

Smart water, grid/power or gas management (what some call the smart* paradigm) constitute massively distributed scenarios that can only be supported with the help of a complex distributed software stack, potentially also requiring the involvement of third-parties, such as telecommunications and cloud operators. For this reason, we must depart from conventional approaches, opting instead for borrowing lessons from the cloud computing and big data domains, in order to cope with the scale and requirements of IoT IACS.

Specifically from a cyber-security standpoint, the distributed nature of modern IACS makes it difficult not only to understand the nature of incidents, but also to assess their progression and threat profile. Reacting and defending against those threats is something that is becoming increasingly difficult, requiring orchestrated and collaborative distributed detection, evaluation and reaction capabilities well beyond the reach of a single entity. Therefore, the current approach of Cyber Security for IACS has to be improved with new tools and models capable to protect the whole value chain of a CI in increasingly sophisticated and networked scenarios.

This poses several challenges that can only be adequately tackled by creating defence and reaction

architectures built from the ground up to support collaboration and information exchange, while making efficient use of such information to assess security threats and anticipate their progression, providing a decision support system for cyber-security. This is one of the most relevant contributions of the ATENA H2020 project (http://atena-h2020.eu). Specifically, this paper will present the ATENA cyber-detection architecture, designed from the ground up to address the particular needs of modern distributed IACS, while providing (near)real-time cyber-security awareness through usage of heterogeneous probes coupled to a distributed big data analysis layer, also encompassing forensic capabilities for post-mortem event analysis and ongoing compliance auditing of security policies. This architecture was designed to provide a considerable degree of flexibility in terms of IACS security management, monitoring and configuration, while preventing risks, both from operational errors and from cyber-attacks, intrusions or malware.

The rest of this paper is structured as follows. Section 2 introduces specific technologies that are going to play a relevant role in the proposal. Section 3 presents the ATENA cyber-detection architecture, encompassing its major functional modules and their integration. Section 4 discusses existing projects and related work in the field of ICS and IACS cyber-security. Finally, section 5 presents conclusions insights about future developments.

2. Security for IACS: requirements and technologies

This section analyses a set of identified requirements as well as some of the most relevant technologies which are going to be leveraged in the scope of the ATENA Cyber-Physical IDS (CPIDS), in order to provide advanced component deployment and security event processing. It starts with the analysis of a contemporary SIEM platform, identifying its shortcomings and limitations in the perspective of the emerging generation of IoT IACS, providing a motivation for the following subsections.

2.1 Designing for conventional ICS: the CockpitCI PIDS as a modern SIEM

In many ways, the CockpitCI FP7 project (CockpitCI 2012) can be considered the predecessor of the ATENA project. Its main objective was to make it possible for a community of CI owners (with interdependent infrastructures) to exchange information about security events. For this purpose, it incorporated a real-time Distributed Perimeter Intrusion Detection System (PIDS) within each CI (see Figure 1) (Cruz et. al. 2014).

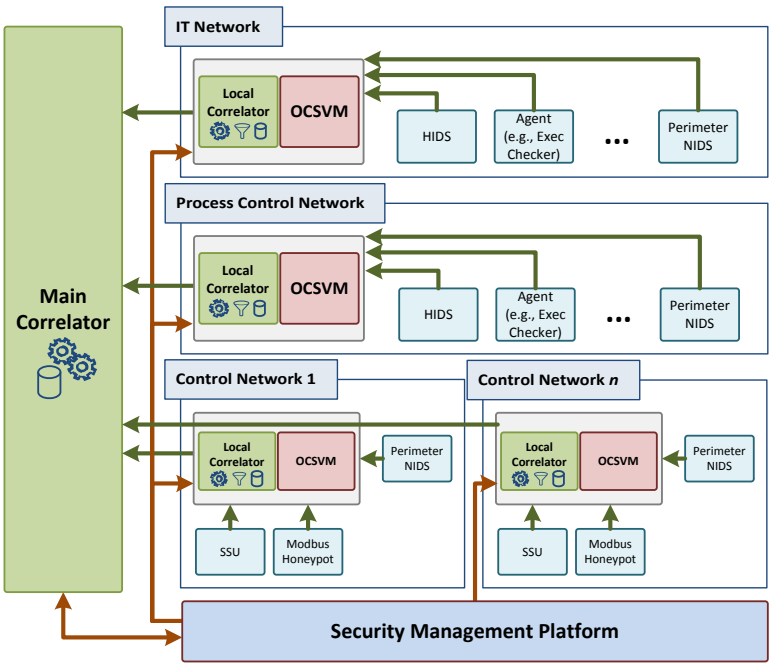

Figure 1: The CockpitCI PIDS architecture.

The CockpitCI PIDS was state-of-the-art for its time, following the same reference model as many contemporary Security Information and Event Management (SIEM) tools. It performed many of the tasks traditionally associated with a Distributed Intrusion Detection System, with support for diversified and closely integrated detection and analysis techniques and tools. Each PIDS was to be deployed in the targeted area of a

CI, in order to detect coordinated cyber-attacks as well as to deploy prevention strategies of isolation, being able to collect and aggregate evidence feeds from probes deployed across the infrastructure, correlating all data in order to detect potential cyber-attacks against systems used to support the operation of CIs.

However, the fundamental premise to the CockpitCI PIDS architecture was based on the idea that CIs were geographically more contained and less distributed than they are becoming today, mostly confined within the space of a production unit (such as an industrial plant) or spread within a single-scope, homogeneous distributed domain. Moreover, its analysis mechanisms were somehow limited and mostly based on signature-based correlation techniques, with the notable exception the One-Class Support Vector Machine (Maglaras et. al. 2014) anomaly detection module for network flows – and more importantly, not suited for distributed deployments and processing of data flows and the rates required for the IoT generation of IACS.

Moreover, probes are statically deployed, sometimes requiring specific configurations that have to be performed via a network management system or, in worst-case scenarios, via direct configuration of the involved devices. This limits the ability to provision and deploy security agents in an on-demand fashion, making it difficult to implement any sort of load-balancing capabilities for evidence collection and processing – an important requirement due to the distributed nature of IoT IACS.

2.2 Event processing using a Lambda architecture

The emergence of distributed IoT IACS architectures brings a paradigm shift in terms of the sheer amount of data being generated and processed. In this perspective, the traditional monolithic cyber detection approaches are not able to cope with the associated scale requirements, while keeping acceptable response times. IoT IACS, involve different domains (physical, logical) and network scopes, also including different types of information sources. How to timely process and analyse all this data, while still accounting for the cyber-physical domain boundaries and interdependencies constitutes one of the main challenges. Moreover, the ever-increasing number of involved components mean that data source heterogeneity and diversity cannot be ignored, as different structured and unstructured sources need to be normalized, stored and processed in an efficient way, while keeping up with (near) real-time turnaround requirements. In this perspective, it becomes obvious why IoT IACS and Big Data are complementary paradigms.

Any sort of cyber detection layer must be designed to provide insights and alerts about the security status of a protected infrastructure. Its operation model is akin to a distributed heterogeneous IDS (DHIDS) architecture, designed to acquire information from several different probes scattered around the infrastructure, which provide evidence about the security status of the protected IACS. These probes include components such as Network IDS, Host IDS, specialized-domain components, and instrumentation data sources or network capture mechanisms providing the information feed that is going to be analysed in order to detect security issues.

From a cyber-physical security (and even safety) perspective, the ability to transport and process security data feeds from the detection agents in a (near)real-time fashion is a critical requirement, as well as the ability to analyse and correlate the information from multiple domains over larger periods of time. The latter case is particularly relevant in the case of slow paced and multi-staged attacks, such as Advanced Persistent Threats (APT)s, which may only be detected using a deeper analysis executed over larger time frames.

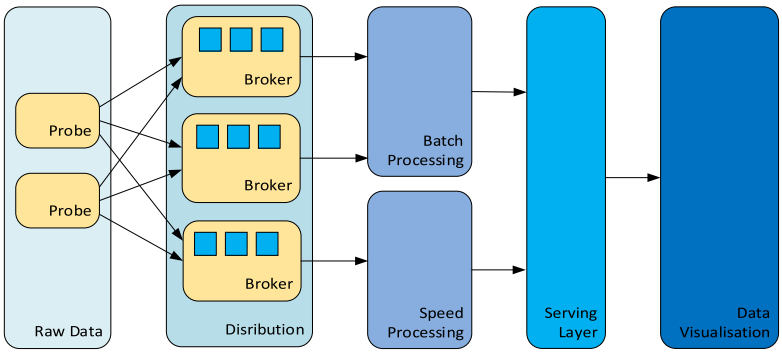

Figure 2: Simplified lambda architecture for the Detection Layer.

Considering the specific and differentiated requirements for event analysis, the authors decided to adopt a simplified lambda architecture (Marz 2015) pattern (see Figure 2), in alternative to a kappa architecture [Forgear 2015]. This allows to accommodate both the needs for quick, as-fast-as-possible (or near real-time) event processing (for critical alerts requiring low reporting latency) and also slow-rate processing (to detect anomalous trends in big data sets). Although it implies maintaining two different pieces of code that are going to coexist in the serving layer, both types of processing have different requirements and it is assumed the need for two different pipelines, as opposed to the kappa architecture where the batch layer is omitted and everything is handled as a continuous feed of data.

The lambda architecture was designed to deal with immutable data sets that grow over time (which is the nature of the security events being generated from the probes). In this line, the fundamental detection layer philosophy implements the best of two worlds: stream techniques for fast, time-window based event processing and batch processing techniques, which constitute a slow path for event processing, sifting through large volumes of data (stored in a large repository, such as a data lake) to search for trends or anomalous patterns. The concept of data lake (not shown in the previous figure) is also relevant in the perspective of the lambda architecture as it provides a persistence mechanism to retain information needed for batch processing, as well as the results of the SIEM analysis paths (both stream and batch). This approach is considered particularly suited to deal with the amount of information generated by IoT IACS, which are ideal candidates for the adoption of big-data event processing techniques, due to their scale.

2.3 Software-Defined Networking and Network Function Virtualization techniques

Using Software-Defined Networking (SDN) and Network Function Virtualization (NFV) capabilities, service support for the components of the security architecture can become flexible and more manageable. SDN is a recent development in network virtualization that allows for the control and data planes to be separated (Kreutz 2015), providing the means to manage and control network flows in a granular fashion. Complementarily, NFV allows for network functions to be separated from the hardware that is running it (Mijumbi 2016).

When used together, SDN and NFV provide a flexible network and function fabric to support the computational and network requirements of the proposed architecture, easing the deployment and management of some of its components. It can be used in specific cases where it is needed to spin up multiple instances of a component, and to rapidly change the network flows between components.

This approach enables the possibility of certain probes to be implemented in the form of virtualized appliances (Virtual Network Functions or VNFs), which can be deployed using SDN mechanisms, allowing for fast and flexible reconfiguration. One of such examples (see Figure 3) relates to the deployment of Network IDS (NIDS) agents: such probes are designed to monitor network communications in a segment of the network. Traditionally, this is achieved by using a monitoring port in a switch, which is configured to mirror all the traffic from the remaining ports of the switch.

The traditional NIDS deployment strategy has some shortcomings. One of the main issues has to do with contention: within a network switch, it is possible for a mirror port not to be able to forward all the incoming traffic of the monitored ports, when these are near their maximum rate. Although some control systems are known to have low throughput traffic, this can vary depending on the specific process, its dimension, and how the network is configured – even if a process generates a low throughput in normal conditions, the figures can rise in the event of a cyber-attack or anomalous situation. In these situations, the switch will drop packets instead of forwarding them to the mirror port where the NIDS is deployed, eventually loosing valuable information.

The use of SDN and NFV can potentially reduce the NIDS contention problem. This is possible by implementing NIDS instances as VNFs, deployed out of the critical control path for the monitored equipment (a crucial requirement for CI applications (Cruz et. al. 2015), using SDN-capable equipment (such as SDN switches) to provide dynamic redirection of monitored flows to the NIDS component. When the maximum bandwidth of the monitoring port is achieved, a new NIDS instance can be spawn on-demand, using SDN to set up new flows to redirect part of the switch communication to the new VNF. Figure 3 illustrates this use case of service support.

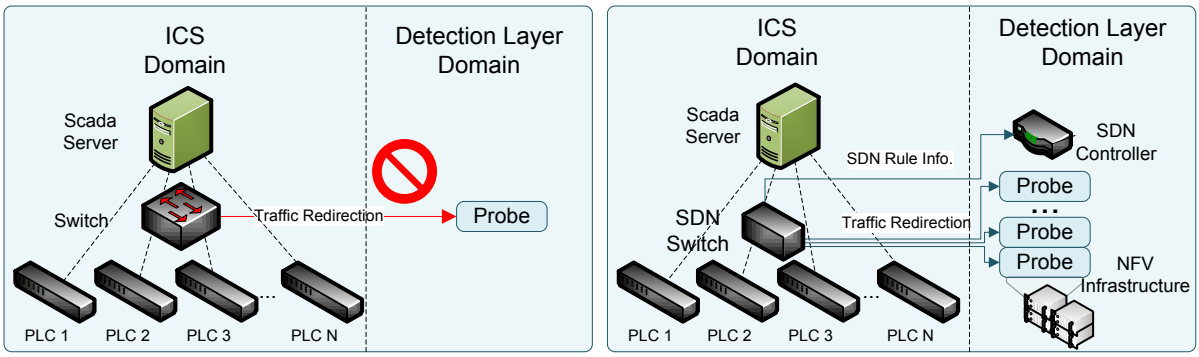

Figure 3: Example using SDN and NFV for service support.

However, the use of these technologies is not a requirement of the architecture itself, but a deployment option. Given the nature of the safety and resilience requirements of the scenarios where this architecture is meant to be deployed, this could not be always possible due to certifications or standards compliance reasons.

3. Proposed Architecture

The ATENA cyber-security architecture, illustrated in the Figure 4 includes several components, namely: different types of probes, from conventional network and host components, to IACS field-specific ones; a Domain Processor per scope, backed by a Message Queuing (MQ) system; a distributed SIEM, for support of streaming and batch processing; a Data Lake, where all the data is stored; and, finally, a Forensics and Compliance Auditing (FCA) module, to enable a post-mortem analysis of the incidents or ongoing compliance validation of organizational security policies. Each of these modules is built on a distributed architecture, designed to accommodate and scale in/out according to the specific needs of the protected IACS (i.e. number events, sources, multiple domains).

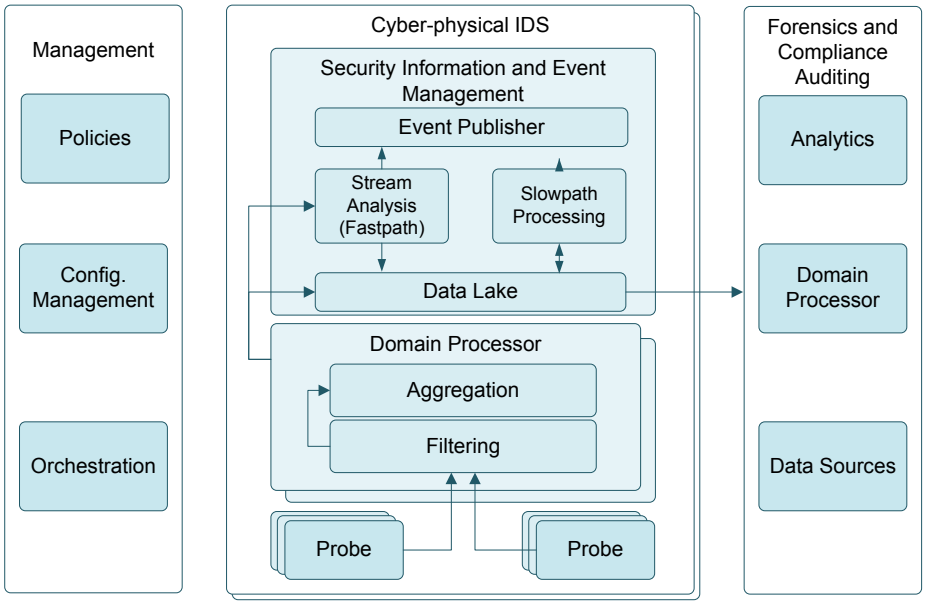

Figure 4: Architecture of the ATENA cyber-security platform.

By following a set of Big Data principles, the main goal of this architecture is to provide an efficient and scalable approach to transmit and process structured and semi-structured data coming from different types of probes to detect and report cyber-attacks and how to converge that information and make it accessible to the upper layers and FCA subsystem.

3.1 Data Streaming Platform

The CPIDS Data Streaming subsystem comprises two main components: the infrastructure to support inter-module connectivity and the **Domain Processors** (one per logical domain in the deployment). The first one

implies a hybrid approach between queuing and publish-subscribe patterns that allows not only for data to be injected by several probes but also to be consumed and processed by several components, in parallel. This constitutes a distributed streaming platform with scale-out capabilities, which can be geographically dispersed, while ensuring, fault tolerance with active redundant replicas, low latency and near real-time performance, by using concurrent approaches and load balancing at the application layer.

The information feeds from the probes are pre-processed near to the collection points, in order to reduce noise and aggregate events before being sent to analysis. **Domain Processors** are responsible for this task, being able to enrich (e.g, via timestamping), filter, count, aggregate, normalize, order and even route events. Despite its capabilities, the Domain Processor is not a full-featured analisys component, rather providing the means to reduce and mitigate some of the data streaming noise. Ideally, the operations performed at this level should not be time-consuming or computationally intensive, under risk of compromising the latency or scalability of the platform.

Regardless of the nature of the events (they may represent from cyber-physical incidents to simple telemetry updates), the streaming platform is agnostic for both the event schema and encoding,. The events are generated by the probes and grouped by topics of interest that are distributed over the entire streaming platform, organized as logical flows of information coming from a set of probes and data sources, to be transported up to the analytics module. Also, multiple topics (the MQ will provide the basic event routing, reliability and ordering guarantees) are used to implement differentiated transport latency and pre-processing requirements for event transport.

3.2 Big data SIEM

The Big Data SIEM, which constitutes the main analytics component of the CPIDS, comprises two types of data modules: streaming (fast path) and batch processing (slow path). Fast path processing is focused on near real time analysis that may be implemented using stream processing of even micro-batch techniques (using micro time-windows). By having more optimized and robust algorithms and more dedicated computational resources, the role of the fast path is to extract meaningful information from the data coming directly from the streaming platform by means of inter-domain correlation, root-cause analysis, inference and or other formal logic methods. Thus, a potential large number of events is transformed into a small number of outputs to the upper-layers, containing a more unified and complete view of the entire system. Low latency techniques, such as lossless filtering, aggregations and stream-based approaches have priority over complex algorithms.

The slow-path (or batch processing) is concerned with large time windows. All the more resource-expensive algorithms, including statistical and machine learning techniques should be part of this path. The idea is to perform deep inspection and anomaly detection over multiple datasets of persisted data (stored in the DL) to find suspicious incidents.

Thus, the CPIDS SIEM relies on Big Data principles and distributed approaches to provide timely results. Both the slow and fast-path processing modules are supported by a distributed data repository (the **Data Lake**) used to collect and retain evidence both from probe-sourced events and outputs produced by the analysis boxes. The DL also provides support for the implementation of a detection layer black box, which can be used to collect and retain evidence, later to be used for forensics or auditing purposes. Finally, a data visualization module will provide the means to monitor and manage the operation of the entire framework, providing a dashboard with information about the ongoing status of the platform.

3.3 Probes

Probes represent the lowest level of the CPIDS architecture, providing the detection capabilities, collecting evidence and providing event feeds regarding suspicious activities. Events are generated using an established format and associated data models, which means that third-party data sources can be converted into probes, by means of adaptors, whose purpose is to normalize data feeds. Among the existing probes, some of the most relevant will be next described:

- **Network IDS:** the perimeter for each network scope (IACS, IT, Access Network, among others) can be monitored using NIDS components, which analyse network flows to search for anomalous or suspicious evidence. These probes have dedicated interfaces to report the security events to the domain processors.

- **Host IDS:** HIDS are deployed in hosts/servers within the system. It is capable of reporting anomalous behaviour in the host where it is deployed.
- **Honeypots:** Acting as decoys in order to detect attackers probing the networks, honeypots provide another source of data for correlation. Besides traditional IT honeypots, the architecture also includes specialized implementations, designed for SCADA networks (Simões et al. 2013).
- **Shadow Security Unit:** The Shadow Security Unit, first developed and proposed in the scope of CockpitCI project (Cruz et. al. 2015), is a low-cost device deployed in parallel with PLC or Remote Terminal Units (RTU), being able of transparently intercept its communication control channels and physical process I/O lines. The SSU provides a cyber-physical security and safety mechanism that requires minimal changes to the existing control networks, being able to work in standalone or integrated within the ICS protection framework. The SSU also implements Machine Learning techniques, based on well-known anomaly detection algorithms, selected accordingly to the computation power of the device, to implement embedded analysis capabilities.
- **Statistical and instrumentation probes:** besides RMON (Stallings 1998), SNMP (Stallings 1998) or Netflow (Claise 2004) sources, the CPIDS can also resort to SDN-related components in some scopes of the CI, namely at the network edge. These sources allow for developing a type of probe, based on the statistical data, provided by network equipment (such as switches) or SDN controllers. These sources can be particularly useful in IoT IACS, for anticipation of threats at the last mile, and those coming from the end users' premises.

All probes publish their event feeds via the Domain Processors. Also, a separate channel (another interface or secure channel, ideally out-of-band) is used for CPIDS management purposes. Eventing channels can optionally be configured to operate within Evaluation Assurance Level 7 (EAL7) (CC 2012) requirements, for added security using unidirectional communications.

3.4 Forensics and Compliance Auditing

The forensics and compliance auditing (FCA) module of the ATENA cyber-security architecture was designed to record and persist digital evidence retrieved from the cyber-analysis layer as well as other sources (such as service logs, Authentication Authorization and Accounting (AAA) sessions or physical access control systems, among others), both for forensics and compliance auditing purposes. While the forensics role provides the means to identify, extract, preserve, and highlight digital evidences for legal (Rani and Geethakumari 2015) or post-mortem analysis purposes, compliance auditing evaluates if standards and policies are effectively met (for instance, regarding authorization procedures for physical installation access, such as access to doors).

FCA procedures have particular interest in the context of CIs (Afzaal et al. 2012), allowing to collect relevant information in the case of a security incident, which can be useful in preparation of response actions or for root cause analysis. Also, the introduction of security of CI-specific quality standards by regulators, service level agreements (SLA) and organizational policies needs to be verified by means of adequate mechanisms. However, existing solutions and procedures cannot be simply applied to CIs due to the latency they introduce (Eden, Burnap, et al. 2015) (Eden, Blyth, et al. 2016).

The FCA subsystem architecture was designed to scale, providing the means to undergo digital forensics activities within a distributed and elastic framework, pretty much in line with the requirements for cloud computing forensics (Morioka and Sharbaf 2015) (Simmons and Chi 2012). The proposed architecture is aligned with a four-stage forensics process workflow broadly defined into four steps, namely: identification, preservation, analysis, and presentation. Departing from the traditional dead forensics model, mostly applied to static data, the FCA subsystem goes for a live systems approach, requiring the implementation of parallel processing forensic engines to allow flexibility in customizing activities for the analysis of evidential data (Hunt and Slay 2010).

From this perspective, the FCA module provides a set of analysis capabilities to retrieve insights from persisted data. Two kinds of actors interact with this architecture: Operators and Security Analysts. Operators receive continuous information from the processes performing rule assessment, evaluating the critically of events and preparing a set of responsive actions to minimize their impact. Security Analysts are responsible for extracting insights from the interpretation of stored events, performing ad hoc queries to understand related thread event paths and preparing improvement measures.

3.4.1 FCA Subsystem Architecture

The core FCA functions encompass different stages, collecting heterogeneous data from internal and external sources, containing structured and unstructured data, to be gathered in a unified view for auditing compliance with a set of rules, also providing forensic investigation functionalities to retrieve evidence. Figure 5 depicts the main blocks, namely the Data Acquisition and Domain Processor (DADP), Data Lake (DL), Monitoring, Orchestration, and Analytics modules, next presented.

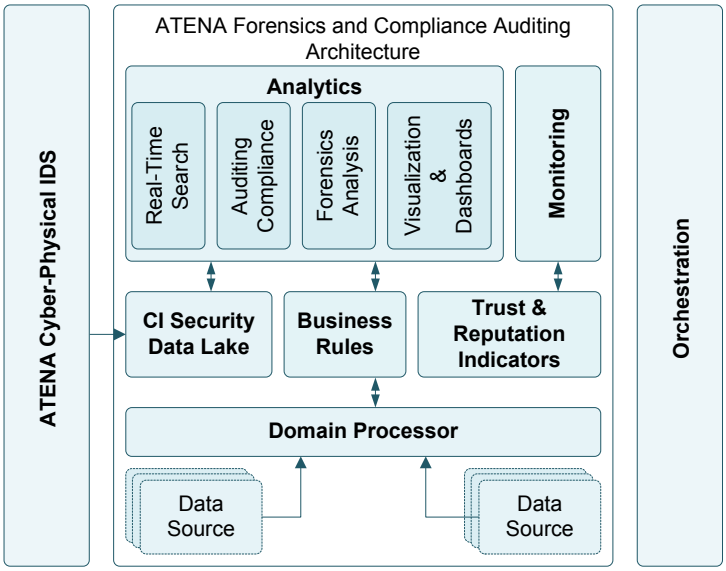

Figure 5: Architecture of the Forensics and Compliance Auditing subsystem.

The **Data Acquisition and Domain Processor** module (DADPM) retrieves information from the environment. It includes a workflow process for management of the incoming flow of "raw data in motion" from disparate heterogeneous structured and unstructured sources, which are normalized into a common format, suitable to be persisted into a **Data Lake (DL)**, for analysis purposes. Moreover, the DL also receives data from the CPIDS and FCA analysis results

The DADPM receives a large amount of data, which is available from data sources within the organizational environment, such as applications, AAA, ICT Security Logs (e.g., from anti-virus), Internal Personnel Activities, Physical Access Control Logs (door switches and surveillance cameras), Maintenance Activities (physical and logical systems), Interactions with third-parties (e.g., general documents, emails) and Incident Logs (e.g. ICT trouble tickets). Integration with third party sources is implemented using custom data adapter components.

The DADPM component also performs enrichment (adding GeoIP data, for instance), filtering, aggregation and indexing of event feeds. Regarding the latter feature, it has the purpose of indexing incoming events in order to enable its efficient exploration by visualization tools for investigation purposes. It may involve creating a keyword index or extract all metadata to be queried later.

A set of **CI Business Rules** provides the knowledge base used for compliance assessment on incoming events. The system is able to provide custom configuration to prioritize or to score different types of alerts (events not complying with the defined rules). For instance, assuming that company policies limit users to using computers within their own organizational unit, any login attempt violating this rule should be reported. Physical access control is another example: alerts could be triggered in when the doors in a given department or physical installation are opened out of a specified period.

Monitoring provides the capabilities to look at things as they happen, allowing operators to look up for abnormalities. It provides visualization information through a **Dashboard** containing information about the number and classification of events through the ingesting workflow and persisted at the DL. Additionally, it is able to display triggered alerts and highlighted information from a continuous assessment of the matching business and audit compliance rules over the events stored in the DL.

Trust and Reputation indicators provide risk levels, allowing to anticipate future services losses or even degradation of service. This information is produced for future analysis, allowing adjustments over security measures and reduce the probability of service failure over dependent services and therefore, to minimize cascading effects that might take place (Caldeira 2013; Caldeira et al. 2013).

The **Analytics** module provides the ability to retrieve insights from a large amount of available normalized events and logs stored in DL. This module provides Data Visualization, Real Time Search, Forensic Analysis and Auditing Compliance functions. **Dashboards** represent the artefacts suitable for visualization of large amounts of data with contextual information, such as the total number and variety of those events or histograms highlighting prioritized events. The **Real Time Search** component of the Analytics module provides rapid search capabilities for information stored on the DL allowing to lookup for data and associated metadata, previously extracted and indexed.

The **Forensic Analysis** and **Audit Compliance** components provide the core functions to apply analytic models in tracing the security activity path of events and identifying non-conforming situations from the correlation of all available data. This may include tasks such as computing security scores, or quantifying the risk levels about the entities involved, also assessing compliance with security standards and business rules.
Finally, the **Orchestration** component takes care of the overall cyber-security framework management, in the same way as it already does for the CPIDS subsystem.

4. Related Work

IACS, and particularly IoT scenarios, are ideal candidates to leverage the benefits of the big data paradigm, both from an operational and also a security perspective. As an example, Guo et al. (Guo et al. 2016) describe an architecture to monitor a large-scale Electric Power System in South China that allows better fault diagnosis, state and forecast estimations. It is also referred how quickly data can grow given the reduced sampling time intervals and the number of inspection points. Cheng et al. (Cheng et al. 2015) describe a city-wise testbed with more than 15 000 sensors whose data processing requirements could not be satisfied using a conventional approach. Finally, Paladini et al. (Paladini et al. 2016) used the same principles to store temporal-series of IoT data from a Smart Building.

Cyber-security detection mechanisms may also benefit from such Big Data-assisted techniques. Las-Casas et al. (Las-Casas et al. 2016), describe an use case that addresses the problem of using these approaches to cope with volume and data source heterogeneity issues, by demonstrating a performance improvement in email phishing counting, by combining Natural Language Processing (NLP) and Locality-Sensitive Hashing (LSH) with big data frameworks. Jia (Jia 2017) discuss the need for deeper analysis based on Big Data processing capabilities to cope with high concealment and long term APT attacks, which are often characterized by weak evidences crossing several layers. Bachupally et al. (Bachupally et al. 2016) present a big data approach to extract features from network data and count specific fields observed in network scans, allowing to efficiently query large amounts of data, as this type of attacks may be concealed over large periods of time.

Kiran et al. (Kiran et al. 2015) used the lambda architecture pattern to process data from ESnet routers for on-line data streaming visualization and job scheduling. Similar, Casas et al. (Casas et al. 2016) refer to a research initiative aiming to combine big data analytics with network traffic monitoring and analysis for both online and offline analysis. All the approaches are in line with the proposed cyber-security architecture, IACS requirements and ATENA project goals, where existent and novel techniques can be integrated to build an efficient, effective and scalable cyber-detection solution.

The forensic capabilities of the proposed architecture come in line with the ideas expressed by Roussev and Richard (Roussev and Richard 2004), which discussed the importance of distributed approaches in forensics and proposed a tool supported by those approaches, where data was centrally maintained and the distributed processing was performed by multiple machines, proving its effectiveness. Also, Ayers (Ayers 2009) proposed several metrics for measuring the performance and efficacy such as speed, accuracy, completeness, reliability, and auditability for measuring the efficacy and performance of forensic tools, claiming that current forensic tools are not keeping pace dealing with the increasing data volume and complexity.
Some of the features of the forensics module were also inspired by other works, such as Alink et al. (Alink et al. 2006), that provide support searching mechanisms in a central database and also allow to store the output

from distinct forensic tools. Bhoedjang et al. (Bhoedjang et al. 2012) later described and highlighted the intrinsic complexity in importing large amount of data with strong query capabilities over a uniform representation. They also noted that new iterations of import tools could include geographic tags. Van Baar et al.(Van Baar, Beek, and Eijk 2014) reported a set of issues from the included cloud features from their approach in the latest iteration of this system.

The use of forensics-oriented visualization components was also suggested by Osborne and Turnbull (Osborne and Turnbull 2009) that highlighted the importance in considering architectures incorporating familiar visualization tools and algorithms that could be able to include distinct data sources, normalizing and correlating its data to retrieve additional insights.

5. Conclusion

This paper presented the architecture for the ATENA cyber-security framework, which is presently under development by a team that includes the authors of this paper, proposes an architecture conceived to address the specific cyber-security needs of the emerging generation of IoT IACS, such as smart grids or large scale distributed SCADA systems.

Leveraging a distributed approach that joins a set of heterogeneous agents and tools together with a big data-based lambda architecture for security event processing, the CPIDS was designed to cope with the scale requirements of protected infrastructures, while providing a coherent framework with flexible provisioning, deployment and orchestration characteristics. Future work will focus on improving the integration with SDN orchestration, diversity of detection agents and automation of provisioning and deployment of components.

Acknowledgements

This work was partially funded by the ATENA H2020 Project (H2020-DS-2015-1 Project 700581).

References

Afzaal, M. et al., 2012. A Resilient Architecture for Forensic Storage of Events in Critical Infrastructures. In *2012 IEEE 14th International Symposium on High-Assurance Systems Engineering*. pp. 48–55.

Ayers, D., 2009. A second generation computer forensic analysis system. *Digital Investigation*, 6, Supplement, pp.S34–S42. Available at: http://www.sciencedirect.com/science/article/pii/S1742287609000371.

Alink, W. et al., 2006. {XIRAF} – XML-based indexing and querying for digital forensics. *Digital Investigation*, 3, Supplement, pp.50–58. Available at: http://www.sciencedirect.com/science/article/pii/S1742287606000776.

ATENA Project Consortium, 2016. ATENA Project Web Site. Available at: https://www.atena-h2020.eu.

Bachupally, Y.R., Yuan, X. and Roy, K., 2016, March. Network security analysis using Big Data technology. In IEEE SoutheastCon, 2016 (pp. 1-4).

Bhoedjang RAF, van Ballegooij AR, van Beek HMA, van Schie JC, Dillema FW, van Baar RB, et al., 2012. Engineering an online computer forensic service. Digit Investig. pp. 96–108.

Caldeira, F.M.S., 2013. *Trust and reputation for critical infrastructure protection*. University of Coimbra.

Casas, P., D'Alconzo, A., Zseby, T. and Mellia, M., 2016, August. Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis. In Proceedings of the 2016 ACM workshop on Fostering Latin-American Research in Data Communication Networks (pp. 1-3)..

CC, 2012. Common Criteria for Information Technology Security Evaluation - Part 3: Security assurance components.

Cheng, B., Longo, S., Cirillo, F., Bauer, M. and Kovacs, E., 2015, June. Building a big data platform for smart cities: Experience and lessons from santander. In Big Data, 2015 IEEE International Congress on (pp. 592-599)..

Claise, B., 2004. Cisco Systems NetFlow Services Export v. 9. Available at: http://www.rfc-editor.org/rfc/ rfc3954.txt.

CockpitCI, 2012. CockpitCI Project. Available at: http://cordis.europa.eu/project/rcn/102078_en.html.

Cruz, T. et al., 2014. Improving cyber-security awareness on industrial control systems: The CockpitCI approach. In 13th European Conference on Cyber Warfare and Security ECCWS-2014 The University of Piraeus Piraeus, Greece. p. 59.

Cruz, T. et al., 2015. Improving network security monitoring for industrial control systems. In 2015 IFIP/IEEE International Symposium on Integrated Network Management. pp. 878–881.

Eden, P. et al., 2016. Forensic Readiness for SCADA/ICS Incident Response. In *Proceedings of the 4th International*

Symposium for ICS & SCADA Cyber Security Research 2016. ICS-CSR '16. Swindon, UK: BCS Learning & Development Ltd., pp. 1–9. Available at: https://doi.org/10.14236/ewic/ICS2016.16.

Eden, P. et al., 2015. A Forensic Taxonomy of SCADA Systems and Approach to Incident Response. In *Proceedings of the 3rd International Symposium for ICS & SCADA Cyber Security Research*. ICS-CSR '15. Swindon, UK: BCS Learning & Development Ltd., pp. 42–51. Available at: https://doi.org/10.14236/ewic/ICS2015.5.

Edwards, C.I.P.M., 2014. An analysis of a cyberattack on a nuclear plant: The stuxnet worm. Critical Infrastructure Protection, 116, p.59.

Forgear, J., 2015. Data Processing Architectures – Lambda and Kappa. *Ericsson Research Blog*, November 19, 2015. Available at: https://www.ericsson.com/research-blog/data-knowledge/data-processing-architectures-lambda-and-kappa.

Guo, Y., Feng, S., Li, K., Mo, W., Liu, Y. and Wang, Y., 2016, August. Big data processing and analysis platform for condition monitoring of electric power system. In Control (CONTROL), 2016 UKACC 11th International Conference on (pp. 1-6). IEEE.

Hunt, R. & Slay, J., 2010. Achieving critical infrastructure protection through the interaction of computer security and network forensics. In *2010 Eighth International Conference on Privacy, Security and Trust*. pp. 23–30.

Jia, W., 2017, January. Study on Network Information Security Based on Big Data. In Measuring Technology and Mechatronics Automation (ICMTMA), 2017 9th IEEE International Conference on (pp. 408-409)..

Kiran, M., Murphy, P., Monga, I., Dugan, J. and Baveja, S.S., 2015, October. Lambda architecture for cost-effective batch and speed big data processing. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 2785-2792)..

Kreutz, D. et al., 2015. Software-Defined Networking: A Comprehensive Survey. *Proceedings of the IEEE*, 103(1), pp.14–76.

Las-Casas, P.H., Dias, V.S., Meira, W. and Guedes, D., 2016, April. A Big Data architecture for security data and its application to phishing characterization. In Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016 IEEE 2nd International Conference on (pp. 36-41)..

Maglaras, L.A., Jiang, J. & Cruz, T., 2014. Integrated OCSVM mechanism for intrusion detection in SCADA systems. Electronics Letters, 50(25), pp.1935–1936.

Marz, N. & Warren, J., 2015. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems* 1st ed., Greenwich, CT, USA: Manning Publications Co.

Mijumbi, R. et al., 2016. Network Function Virtualization: State-of-the-Art and Research Challenges. *Communications Surveys Tutorials, IEEE*, 18(1), pp.236–262.

Morioka, E. & Sharbaf, M.S., 2015. Cloud Computing: Digital Forensic Solutions. In *2015 12th International Conference on Information Technology - New Generations*. pp. 589–594.

Osborne, G. & Turnbull, B., 2009. Enhancing Computer Forensics Investigation through Visualisation and Data Exploitation. In *2009 International Conference on Availability, Reliability and Security*. pp. 1012–1017.

Paladini, F., Marques, C.R., Wanner, L. and Fröhlich, A.A., 2016. Uma arquitetura de banco de dados distribuído para armazenar séries temporais provenientes de IoT.

Roussev, V. & Iii, G.G.R., 2004. Breaking the performance wall: The case for distributed digital forensics. In *Proceedings of the 2004 Digital Forensics Research Workshop (DFRWS 2004)*.

Simmons, M. & Chi, H., 2012. Designing and Implementing Cloud-based Digital Forensics Hands-on Labs. In *Proceedings of the 2012 Information Security Curriculum Development Conference*. InfoSecCD '12. New York, NY, USA: ACM, pp. 69–74. Available at: http://doi.acm.org/10.1145/2390317.2390329.

Slay, J. & Sitnikova, E., 2009. The Development of a Generic Framework for the Forensic Analysis of SCADA and Process Control Systems. In M. Sorell, ed. *Forensics in Telecommunications, Information and Multimedia: Second International Conference, e-Forensics 2009, Adelaide, Australia, January 19-21, 2009, Revised Selected Papers*. Springer Berlin Heidelberg, pp. 77–82. Available at: http://dx.doi.org/10.1007/978-3-642-02312-5_9.

Stallings, W., 1998. SNMP, SNMPV2, SNMPv3,and RMON 1 and 2, 3rd ed., Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Van Baar, R.B., van Beek, H.M.A. & van Eijk, E.J., 2014. Digital Forensics as a Service: A game changer. *Digital Investigation*, 11, Supplement 1, pp.S54–S62. Available at: http://www.sciencedirect.com/science/article/pii/S1742287614000127.