Hindawi
Security and Communication Networks
Volume 2021, Article ID 9947347, 10 pages
https://doi.org/10.1155/2021/9947347

Review Article

Migrating to Zero Trust Architecture: Reviews and Challenges

Songpon Teerakanok,[1,2,3] **Tetsutaro Uehara**,[1,4] **and Atsuo Inomata**[1,5]

[1]*Cyber Security Laboratory, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*
[2]*Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom 73170, Thailand*
[3]*Research Organization of Science and Technology, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*
[4]*College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan*
[5]*Osaka University, Suita, Japan*

Correspondence should be addressed to Songpon Teerakanok; songpon.te@cysec.cs.ritsumei.ac.jp

Received 10 March 2021; Accepted 7 May 2021; Published 25 May 2021

Guest Editor: Qi Li

Zero trust (ZT) is a new concept involving the provisioning of enterprise/organization resources to the subjects without relying on any implicit trust. Unlike the perimeter-based architecture in which any subject behind the wall (i.e., inside the predefined perimeter) is considered trusted, zero trust architecture (ZTA) processes any request and provides a resource to the subject without relying on implicit trust. In this paper, based on NIST Special Publication SP800-207, the concept of ZT and ZTA is introduced. Also, challenges, steps, and things to consider when migrating from the legacy architecture to ZTA are presented and discussed.

1. Introduction

The advancement of technologies brings forth new and more convenient ways of living through the invention of smarter tools and services. The proliferation of cloud technologies and the Internet of things (IoT) bring revolutionary changes to today's IT systems. These changes, however, also come with great challenges. As the IT systems grow, also hackers and malicious actors' skills are adapted and honed rapidly to an astonishing degree. An example of a recent sophisticated attack in December 2020 on SolarWind Orion is given [1], which is an IT monitoring and management software, affecting global victims. This global intrusion campaign was carried out by using a supply chain attack via a trojan (so-called "SUNBURST") backdoor. Using several obfuscating and evasive techniques, the campaign is believed to be the work of a highly skilled threat actor.

1.1. Perimeter Security. "Everything inside the internal corporate/organization network is considered trusted." Until now, this is the concept that we believed and built our network/system upon. The traditional idea of border protection divides networks into two main areas: internal and external networks. The internal network covers every subject (e.g., devices) within the predefined border based on the devices' physical location while the external network covers the rest. Firewalls, IDS, IPS, and other security controls are usually deployed at the edges (borders) of the corporate/organization network to draw a secure boundary (or "network perimeter") that separates its internal network from the rest of the Internet. This idea forms the basis of perimeter-based network security.

Generally, the legacy method of perimeter-based security allows the use of an implicit trust in which once the subject is authenticated and allowed to enter the internal network, it is considered trusted. This allows a malicious (compromised) subject to perform further lateral movement and roam freely once it infiltrated the internal network [2].

1.2. Zero Trust Concept. As technologies continue to advance, the demands and lifestyles of users are also changing rapidly. Cloud technologies offer us new ways to access services and resources anywhere and anytime with high cost performance. Nowadays, people are no longer required to work from their office/workplace; instead, they can work remotely from anywhere as long as all resources needed to perform their job are available. With the emerging of a Bring Your Own Device (BYOD) policy and the current COVID-19 pandemic situation, remote working and working from home (WFH) have become a common thing (so-called "new normal") for many organizations. For example, in November 2020, Square Enix, a large Japanese video game company, offered their employees an option to permanently work from home [3, 4] which greatly demonstrates a paradigm shift in this regard.

Now, the question lies in how this change in paradigm affects an organization in terms of security. The legacy perimeter-based network security is considered insufficient since the users are currently allowed to work remotely from any place which may no longer be located inside the secured perimeter. Therefore, it has become very difficult to define or draw the exact borders/perimeters, let alone securing them.

These problems brought about the concept of "Zero Trust" (ZT) in which an enterprise must assume that there is no implicit trust in every subject. In the ZT security model, the enterprise-owned environments are considered no more trustworthy than any nonenterprise-owned environment [2]. More details regarding ZT and zero trust architecture (ZTA) are provided in Section 2.

In this paper, we discuss the adoption of the ZT concept and ZTA to the enterprise/organization. Since ZT is an evolving concept which cannot and will not be completed by just buying and replacing all the network equipment, there are many factors an organization needs to consider while migrating from the legacy perimeter-based model to ZTA. In this paper, details of threats and challenges in transitioning from traditional network to ZTA are introduced. Furthermore, some key factors and basic guidelines for ZTA migration are presented and discussed.

The rest of this paper is organized as follows. Section 2 presents a brief introduction to the concept of zero trust (ZT) and zero trust architecture (ZTA). In Section 3, we discuss security requirements needed for ZTA deployment and new threats against ZT-based systems. Section 4 presents and discusses processes and factors needed to be considered to successfully migrate to ZTA. Next, the remaining challenges, details, and steps for ZTA implementation are presented and discussed in Sections 5 and 6, respectively. Finally, we summarize this paper in Section 7.

2. Zero Trust Architecture

Zero trust architecture (ZTA) adopts an idea in which all subjects are implicitly considered untrusted no matter where they are located (either internal or external), which is the opposite of how perimeter security works. However, it does not mean that there is no trust at all in ZTA. In this new way of thinking, typically, a subject earns trust from the system on a particular request/transaction by proving itself through authentication and authorization. Enforcing the authentication and authorization process on every request/transaction gives the system the ability to granularly control and adjusts the security level required to access a particular resource.

In the following subsections, components, common models of ZTA, and a brief introduction to the trust algorithm are presented and discussed, respectively.

2.1. Components. There are 5 major logical components in ZTA as displayed in Figure 1: subject, resource, policy decision point (PDP), policy enforcement point (PEP), and supplement [2]. Subject refers to a user or any device requesting access to the enterprise resources. As the name suggested, resource refers to the corporate/enterprise resource being requested by a subject. The resource can be either single or multiple pieces of resources depending on the content of the request.

A policy decision point (PDP) is responsible for deciding to allow or deny access to the enterprise resource and establish or terminate the communication between a subject and a resource being requested. PDP can be broken down into two components: policy engine (PE) and policy administrator (PA) which are responsible for decision making and communication management, respectively.

A subject sends a request to access an enterprise resource which ends up sending to the policy enforcement point (PEP). The PEP forwards the request to PDP. After the PDP decides what to do with the request, it then issues a command to PEP to enable or terminate the communication between the subject with the resource. As we will see, PEP acts as a gate between the subject and resources. Not only controlling the flow of the communication but PEP is also responsible for monitoring the network traffic going between the subject and the requested resource.

Lastly, supplement helps to provide useful information (e.g., threat intelligence information and network/system logs) to the PE. This allows PE to make more accurate and correct decisions (less false positive and false negative) which also ends up enhancing the overall security of the system.

2.2. ZTA Models. NIST SP 800–207 [2] classifies ZTA deployment into 4 types, i.e., device agent-based, enclave-based, resource portal-based deployments, and lastly the ZTA deployment using device application sandboxing, depending on how resources are managed and safeguarded. There are both agent and agent-less approaches with PEP acting as a gateway to the enterprise resources. The PEP can be attached to a single or multiple resources or act as a portal to all enterprise resources.

Also, there is a special variant of an agent-based model utilizing sandboxing, which could be a virtual machine (VM) or containers such as Dockers. The goal of utilizing a sandbox in this model is to prevent the trusted application from any malicious activities that may originate from a possibly compromised subject (host). Please refer to Section 3.2 of [2] for further information regarding the ZTA deployment model.

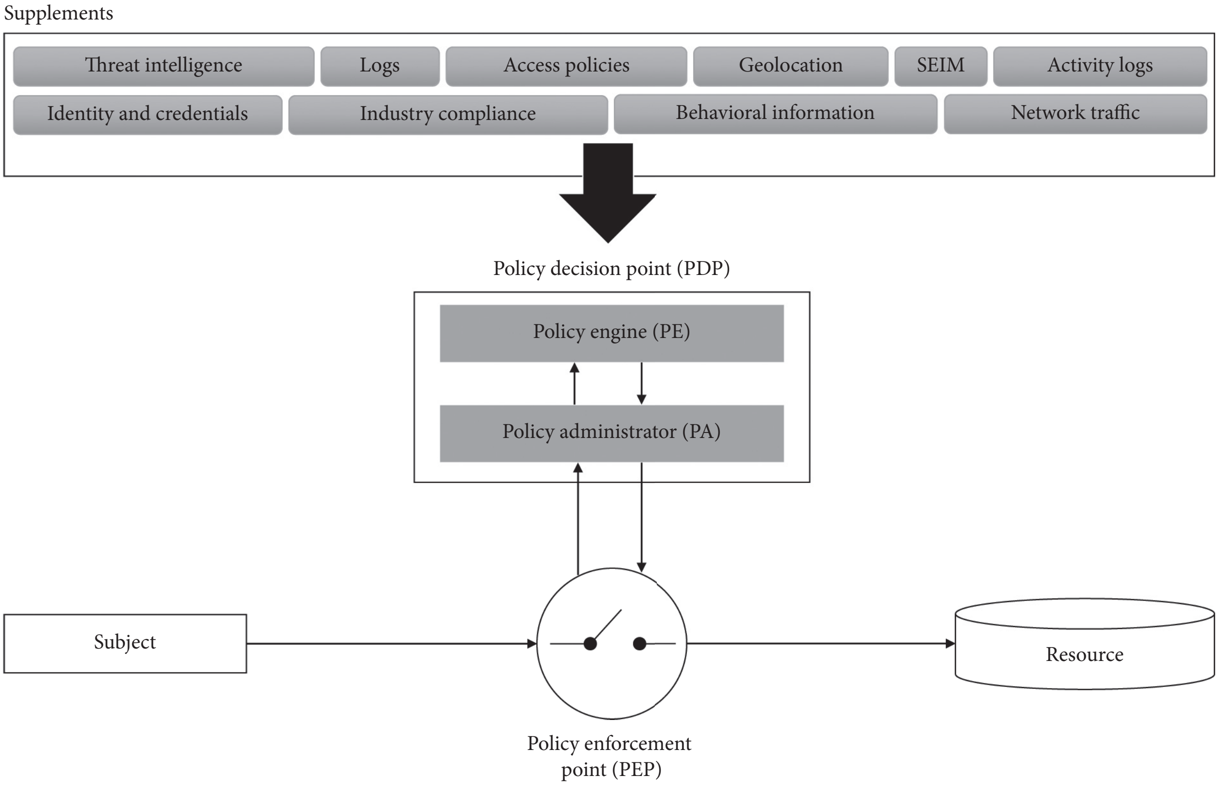

Figure 1: Logical components of the zero trust architecture (ZTA).

2.3. Trust Algorithm. While the policy engine (PE) can be considered a brain of the ZTA, inside this brain holds a crucial thought process known as trust algorithm. The trust algorithm (TA) is used by PE to decide whether to grant or deny access to the requested resource. To draw such a decision, PE generally incorporates several pieces of information from various sources which may include access request (i.e., request content), subject information (corresponding to the subject database), asset database (known enterprise assets including BYOD devices), resource requirements, security and network traffic logs, and threat intelligence information.

There are some trust algorithms proposed in the literature. Chen et al. [5], in 2020, introduced the use of behavior-based anomaly discovery technique incorporated with user and devices identities, user behavior, terminal security status, and system behavior information to perform the trust assessment process. The proposed method utilizes the hierarchical trust level-based access control, in which the resource access is granted only when the trust level of a subject reaches a certain level. This approach offers a dynamic access control with fine-grain tuning capability using the required trust level/threshold.

Similarly, in October 2020, Yao et al. [6] proposed a dynamic access control and authorization system for ZTA. The authors proposed a trust-based access control (TBAC) which is based on the calculating of behavior trust (BT) score for each user. In case that BT exceeds the minimum trust score required to access the resource, access permission is granted. In this method, user behaviors affecting the BT score calculation are login behavior (e.g., login method and time), network behavior (e.g., amount of TCP traffic), and operational behavior (e.g., resource name and access history).

Lastly, Vanickis et al. [7], in 2018, proposed a policy enforcement framework for the zero trust network (ZTN). In this paper, FURZE, a framework for risk adaptive access control (RAdAC) is proposed. Using the RAdAC-based approach allows the system to utilize both operational need and security risk to accurately and dynamically grant or deny access to the enterprise resource. Furthermore, the authors introduce two domain-specific policy languages PAROLE (similar to XACML [8]) for configuring access control to network resources and FACL (firewall access control list) which is designed for expressing firewall-specific filtering rules and configurations.

3. Security in ZTA Environment

New technology comes at a security cost. Cyberattacks are continuously being refined and becoming more sophisticated. To adopt the zero trust concept to the organization, there are things to look out for and to be aware of. In this section, threats and challenges that might come with ZTA are introduced and discussed. Furthermore, we discuss some requirements needed to successfully deploy ZTA in practice.

3.1. Threats. Most of today's information system has its soft spots whether they are technical (e.g., system designs and configurations) or nontechnical (e.g., human). Taking advantage of these weaknesses, adversaries prey on an enterprise by exploiting these vulnerabilities. To defend against such attacks, the organization should first get a clear grasp of what the attack surface looks like. Understanding the attack surface allows us to have insights on where (in which way and how) the attack might be carried out. In the following subsections, we discuss the new attack surface and some threats associated with ZTA.

3.1.1. New Attack Surface. As shown in Figure 1, the policy decision point (PDP) and policy enforcement point (PEP) are the core parts of ZTA. It is responsible for making any decision whether to grant or deny access to the enterprise resource and also managing the connection between the subject and the resource. In ZTA, these core parts (i.e., PDP and PEP) can be new targets for adversaries. Unlike the traditional perimeter-based architecture, ZTA may suffer from attacks on these core parts such as DDoS, route hijacking, or supply chain attacks on PDP. By disrupting PDP and PEP operations, the operations of the enterprise network may come to a halt. Furthermore, in case of PE being compromised, it may cause serious harm (e.g., data tampering, or leakage) to the organization.

Furthermore, since there is no longer a wall to protect enterprise assets, all assets have the possibility of being targeted. Compromised assets or accounts with a high level of access privilege, especially ones with access permission to the resources interested by an attacker (e.g., commercial, and financial information), are likely to be primary targets of attacks.

3.1.2. Denial-of-Service (DoS) and Network Disruption against PDP. As mentioned earlier, the PDP, consisting of policy engine (PE) and policy administrator (PA), can be a new target of attack since every decision to grant or deny access to the resource is determined by PDP. Corrupting or disrupting PDP will also greatly affect such decision process which can result in a halt in operations. One way to mitigate such network disruption attacks is to put PDP (i.e., PE and PA) on the secured cloud environment which is more resilient in the face of such an attack.

3.1.3. Unauthorized/Unapproved Changes in PDP. According to NIST SP800-207, the system administrator may perform unauthorized changes or accidentally creates misconfigurations which may disrupt or create vulnerabilities in the system. A misconfigured or compromised PE might grant access to some restricted resources that would otherwise not be allowed. Also, a compromising PA may allow adversaries to bypass PE's decision process and access the enterprise resource directly.

Mitigating suck risks, as suggested in NIST SP800-207 [2], involves the logging and monitoring of PDP activities. Moreover, PE and PA should be properly configured with all changes being documented. Lastly, both PE and PA should be subjected to audit.

3.1.4. Credential Theft. In ZTA, all assets are not implicitly trusted and have the possibility of being attacked based on their importance and the importance of the information they hold. Some assets or accounts are more likely to be targeted. Compromising an account or asset is not something new and is not unique to the zero trust architecture. With the new BYOD policy, it may be even easier for an attacker to successfully compromise a BYOD asset. Since BYOD devices are not controlled by the enterprise, they may not receive the latest security patch or may not have any antimalware mechanism installed. Since there is no such wall of protection as in the perimeter-based architecture, a well-developed ZTA should prevent or hinder compromised assets or accounts from accessing enterprise resources.

One way to mitigate such a problem is to monitor the subject's behavior which might include login history and pattern, duration, and resource access pattern. A subject accessing any resource outside its normal access pattern may raise a flag which can lead to a more thorough investigation. A good example of deploying a behavior-based approach in ZTA is presented by Yao et al. [6] in 2020. In [6], the subject's behavior is continuously observed and calculated into a behavior trust (BT). Access to a resource is granted only if BT exceeds the trust threshold (TT) which may change dynamically depending on the environment.

3.1.5. Network Traffic Monitoring and Inspection. ZTA relies on end-to-end communication which usually contains encrypted information. Some third-party software/services are fully encrypted making it very difficult or impossible to perform full packet inspection. This leaves the enterprise no choice but to perform packet analysis based on the metadata of the packet. However, it is suggested in [2] to incorporate machine learning techniques to help to analyze the encrypted traffic for better efficiency.

3.2. Requirements. There are three primary requirements needed to be fulfilled to successfully deploy and implement ZTA.

3.2.1. Granular Data Visibility, Access Control, and Data Protection. One key point of ZTA that differentiates it from the legacy perimeter-based model is how it protects the resources. ZTA protects individual resources while perimeter-based architecture protects all enterprise resource as a whole. In the case of breaches, perimeter-based architecture tends to suffer greater damage comparing to ZTA. To protect individual resources, enterprises are required to adopt the data-centric approach [9] including data/resource discovery, tracking, and analysis. Enterprises are expected to know what kind of data and resources they are holding, how they are protected, and who and when accessed these resources.

Enterprises should exercise the least privilege policy [10] so that there is only the least amount of data and resources available to a subject. The least privilege policy can be applied with dynamic trust-based access control (see [5, 6], for example) to provide granular visibility of data. Also, an individual resource should be protected at the border of that resource (using resource gateway, for example).

3.2.2. No Implicit Trust. “Guilty until proven innocent” [10, 11] is the term that best describes ZTA. A subject is trusted conditionally (i.e., only when some conditions are met). To be trusted, a subject is required to perform some action (e.g., authentication) to earn it. Nothing in ZTA has implicit trust; the only way to gain trust is to earn through verification.

The verification comes not only in the form of authentication but also in policies and requirements. All subjects are required to meet the minimum security requirements and access policies. For example, the agent, in the agent-based model, may drop the request automatically should the device performing a request is not patched or not meet the minimum security requirements set by the enterprise.

Not only trust can be earned but can also be lost. For example, the PDP may decide to reduce the level of trust given to a subject if it performs some suspicious activities such as requesting a resource outside the subject’s scope of permission or performing multiple login attempts in a very short period. Lastly, trust is not a constant value, and its value should be decreased over time. It means that even though a subject can access a certain resource at the moment, the same right to access this particular resource is not guaranteed in the future. The re-evaluation of a subject’s trust is needed. An enterprise is recommended to apply the decaying property to the trust algorithm in its zero trust system.

3.2.3. Continuous Authentication and Monitoring. One important requirement for building a zero trust system is continuous monitoring and evaluation. In ZTA, a subject earns trust separately for each of its requests. In ZTA, requesting the same resources at different times should require verification and re-evaluation of trust. As mentioned earlier in the previous subsection, trust can be earned and also can be lost. Therefore, continuous monitoring of the subject activities is required. Continuous monitoring helps maintain the current level of trust a subject has while detecting inconsistencies or anomalies in a subject’s behavior (e.g., changes in access pattern and failed login attempts) which are crucial information in re-evaluating trust.

Continuous authentication and monitoring do not mean a user has to type in his/her password for every single request. The process of continuous authentication should be done in a nonobtrusive manner [9] to increase practicability. A subject may perform multifactor authentication (MFA) at first and maintain its trust level by allowing the system to observe and analyze its behavior (e.g., request pattern, keystroke, geolocation, or network traffic).

An early example of continuous monitoring is the work of Brosso et al. [12] in 2010. In [12], a continuous authentication system based on user behavior is proposed. The proposed method utilizes user behavior in determining the level of trust for each user. By incorporating the neuro-fuzzy logic technique to continuously update the user behavioral database, the author introduces the new method to calculate trust with better accuracy.

4. Migrating to ZTA

Deploying and implementing ZTA is a multiple-step and continuous process that cannot be achieved by replacing all tools and equipment with new ones. To introduce the zero trust concept to existing perimeter-based architecture, there are several steps and factors to concern. Figure 2 shows the ZTA deployment cycle which is originally based on the NIST risk management framework (RMF) [2, 13].

In this paper, we divide the migration process into three major steps. The following three subsections explain the details of each step.

4.1. Assessment. Assessment involves several things including identifying the actors and assets of the enterprise. The enterprise should have a clear grasp of its subjects including both human and nonperson entities (NPEs). It should be able to identify and monitor both enterprise-owned and nonenterprise-owned assets, including both hardware and digital artifacts (e.g., software and digital certificates). Furthermore, the enterprise is expected to have the ability to configure, manage, and observe the current state of the asset.

Dealing with “shadow IT” and nonenterprise-owned assets such as BYOD devices is certainly more challenging compared with managing enterprise-owned assets. However, the enterprise should try its best to discover and observe these assets [2]. Lastly, depending on the critical mission of the enterprise, it may also need to list and categorize its high-value assets (HVAs) and all relevant processes.

4.2. Risk Assessment and Prioritization. An enterprise should identify and rank its business process based on the criticality and importance of its mission. By studying the risk impact and performing some prioritization, the enterprise may decide to start migrating its very first business process that has relatively low-risk to ZTA and then continues later with a more critical business workflow/process after experiencing and learning from the first transition.

Once the candidate business process has been selected, now, it is time to create policies for this candidate process. In this step, all resources used or affected by the candidate process should be identified. This allows the enterprise to know precisely what resources are involved with the migration process. Lastly, in the case of a criteria-based trust algorithm, a set of criteria is determined. On the other hand, regarding the score-based trust algorithm, trust/confidence level weights of each resource used in the candidate process are initially defined. Both the mentioned set of criteria and the weights are expected to change during the initial tuning period and may also change over time to ensure its effectiveness and practicality.

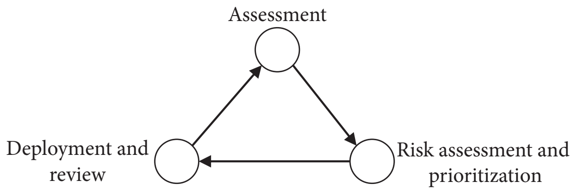

Figure 2: Development cycle of ZTA (based on NIST SP800-207 [2]).

4.3. Deployment and Review. After the enterprise determines the candidate business process to migrate to ZTA, the actual deployment begins. Implementation of a new ZTA-based business workflow must follow the security policies developed in the earlier phases. Generally, the deployment always associates with the logging and monitoring process. Analyzing logged and monitored information (e.g., request pattern, login attempts, and communication patterns) allows the enterprise to ensure that the new ZTA-based business process works effectively as intended.

In this step, results including mistakes during earlier phases are collected. The organization analyzes and learns from mistakes and makes changes (if necessary). The results and changes should be properly documented. With each migration, the enterprise/organization gains more confidence and can choose a more challenging workflow for its next migration cycle.

4.4. Things to Consider during ZTA Migration Process. There are few things that an enterprise needs to pay close attention to during migration to ZTA.

4.4.1. Changes Procedure. When making changes during the migration process, all changes are needed to be properly documented. Since migrating to ZTA is the change from the lowest level of the architecture, i.e., the fundamental concept, therefore, there is a possibility that the enterprise needs to perform several changes (both hardware and software) or even redesigns some of its workflows from scratch. Therefore, a well-designed change management policy and plan are required. Well-documented change procedures help the enterprise keep track of what is changed and help provide useful information once something goes wrong.

4.4.2. Risk Management. Risk management is not something new or unique only to ZTA. Generally, the enterprise is required to identify and assess the risk associated with its missions or business workflows. Deployment and implementation of ZTA may cause the enterprise to reidentify and re-evaluate risks associated with any business process involved. NIST risk management framework, in the Special Publication SP800-37 [13], provides a guideline in this regard which is also applicable with ZTA.

4.4.3. Identity Management. ZTA involves and relies heavily on the subject identity to perform subject provisioning. To grant a resource to a subject, the policy engine needs several pieces of information including identity and credential information to perform a decision. Therefore, the enterprise should also make changes to its ICAM [14] policy to include and support new ZTA-based workflows. NIST provides recommendations regarding digital identity management through its special publication SP800-63-3 (i.e., Digital Identity Guidelines [15]) which are also applicable to ZTA.

4.4.4. Laws and Regulations. To migrate to ZTA, the enterprise should make efforts to ensure that new ZTA workflows comply with laws and regulations, for example, Health Insurance Portability and Accountability Act (HIPAA) [16], Payment Card Industry Data Security Standard (PCI-DSS) [17], and General Data Protection Regulation (GDPR) [18].

5. Challenges in ZTA

To successfully implement a good zero trust system, there are currently many difficulties an enterprise needs to overcome. In this section, we discuss some current problems and some remaining challenges that hinder the deployment of ZTA in practice.

5.1. Vendors Lock-In and Interoperability. The vendor lock-in problem is not something new. We experienced this problem before in both IoT and cloud platforms. In an IoT system, we usually find some of our devices unable to communicate and interoperate with devices from different vendors. For example, some of Google's IoT devices (e.g., Nest Thermostat) are not compatible with Apple's IoT ecosystem (i.e., HomeKit).

Regarding cloud platforms, the vendor lock-in problem refers to the restriction enforced by the cloud provider to prevent or discourage users from switching a service. Generally, cloud providers encourage new users to easily sign up and offer their services at a low initial price. However, once the user decided to scale the service, the price starts to grow exponentially. At this point, the user may consider switching to other cloud providers with better offers. This is also when the vendor lock-in problem happens. A user switching to the other providers means the company losing its revenue; therefore, the cloud provider may try to impose some technical difficulties, legal restrictions, or some additional fees to discourage the user from leaving.

Some providers are utilizing proprietary technologies which also make the migration even harder. Opara-Martins et al. [19] provided a comprehensive analysis highlighting the impact of vendor lock-in problem in business perspective. Surprisingly, the study shows that many customers including decision-makers of many companies lack sufficient knowledge and awareness regarding proprietary technologies that might prevent or restrict interoperability and portability when procuring cloud services from providers.

Concerning vendor lock-in problems in ZTA, zero trust is a new evolving concept which is not fully matured at the moment. Therefore, there is no single-vendor solution available from any vendor at the moment. For many existing enterprises and organizations, migrating to ZTA is a continuous process that might take a long time. Therefore, some organizations may prefer purchasing components from different vendors according to their needs instead of purchasing from any single vendor.

However, to avoid vendor lock-in problems, some standards to support interoperability between devices are needed. As we already know, many vendors usually rely on proprietary APIs rather than the standard. This makes it very difficult for two devices from different vendors to work together smoothly since they follow different protocols and API. Also, in the case of partner companies changing their API behavior, vendors are required to make changes to support them as well.

There is no silver bullet to this problem. The problem of vendor lock-in is simple but hard to solve and will continue to exist. To select suitable technologies, platforms, or infrastructure to support ZTA, an enterprise needs to first identify dependencies in their IT system. If its current IT systems are designed to operate or rely on legacy technologies, there might be left with only limited options to choose from. Furthermore, if the current IT systems are compatible with only limited technologies and platforms, the enterprise may consider upgrading the legacy systems before migrating to ensure the compatibility and interoperability of these IT systems. These upgrades will help to prevent future loss should the company decide to switch from one vendor to another.

5.2. Proprietary Data Formats and Need for Standardization. The decision-making process of ZTA is done by the policy decision point (PDP). This process requires information from various sources to draw the final decision on whether to grant or deny access to the enterprise resource. Although some of them already have standards for exchanging information, for instance, STIX [20] and TAXII [21] for sharing threat intelligence information, there is still no common standard for some of them. In the case of a provider has encountered some security or technical issues and the enterprise needs to find a quick replacement, it is very difficult to do so without paying a high price [2].

5.3. Avoiding User Disruptions. One of the biggest challenges in implementing ZTA is to avoid disrupting users during deployment. At the start of each migration cycle, an enterprise may start the migration process by introducing a new ZTA-based approach, which is designed to replace the legacy system, inside the perimeterized zone and encourage its users to utilize this new ZTA-based system. The enterprise may gradually impose technical restrictions to the old method at the same time to encourage users to switch to the new one. As more users utilize the new system for an extended period without any problem and exception, the enterprise may decide to move the new workflow from the perimeterized zone to the unprivileged network. Finally, the enterprise needs to perform a final clean-up to remove any decommissioned or deprecated service to complete the migration process.

5.4. Trust Level and Resource Classification. Requesting resources in ZTA involves calculating a trust level from credentials and information provided by the subject at the time of request and comparing it to the predetermined trust level required to access the particular resource. Determining an appropriate level of trust for each resource is a challenging task. The enterprise needs to determine an appropriate level of trust which is not too high or too low for each resource. Too high makes the resource too hard to access which may also end up hindering the workflow, while too low means the resource is too easy to access and less secure.

5.5. Dealing with Unmanaged Devices. Some companies, such as Google (see [22–27] and Section 6.4 for further details), decide to implement ZTA by allowing only corporate-owned or managed devices to access corporate resources. However, some companies may exercise a BYOD policy that might not allow them to freely monitor, control, or install any agent software on employee's personal devices due to privacy issues. Therefore, allowing unmanaged devices to securely access corporate resources without imposing too many restrictions is also one of the remaining challenges in deploying ZTA.

5.6. Improving Trust Algorithm. Technically, the trust algorithm (TA) is considered the thought process inside the brain (i.e., PDP) of ZTA providing PE an ability to accurately decide whether to grant or deny access to all incoming requests. TA incorporates information from various sources including threat intelligence, SIEM logs, network traffic, subject's geolocation, user's identities, and credentials.

Each piece of information is not equally important; some information, such as user credentials, are more important and are weighted more, comparing to other factors such as network traffic, in calculating a trust level of a subject. Currently, there is no optimal solution, guideline, or reliable approach in weighting such factors; the enterprise implementing ZTA needs to continuously observe and adjust these parameters over time to ensure it functions accurately as intended.

TA may consist of both static rules (e.g., deny all access from a known compromised or malicious device) and a dynamic decision mechanism that calculates the possibility of a subject being malicious based on information available at the time of the request. This dynamic decision-making approach usually involves the use of machine learning (ML) allowing TA to heuristically improve its decision-making capability over time. As mentioned earlier, TA is considered a crucial part of ZTA. Inaccurate or tampered results from TA can greatly affect PDP's final decision which might end up allowing a malicious subject to access corporate resources (i.e., false positive) or deny legitimate users from accessing the resources they need (i.e., false negative). Therefore, an enterprise is required to put efforts into fine-tuning TA from time to time to maintain and improve its functionality.

Finally, TA incorporates information from many sources which usually come in different formats. Some information is redundant, while some are poor in quality or irrelevant. Therefore, this information is needed to be normalized, filtered, and then correlated to improve the overall efficiency of TA.

6. Implementing ZTA

In this section, we discuss generic details, steps, and information regarding the implementation of ZTA. Every ZTA implementation is a unique and completely different journey for each organization; therefore, we will only discuss generic details and common issues in putting ZTA into practice. There are 3 major steps in implementing ZTA. Sections 6.1–6.3 explain these steps in detail. Lastly, in Section 6.4, we briefly discuss a real case study of ZTA implemented by Google, called "BeyondCorp."

6.1. Identify Devices and Users. ZTA involves heavily in managing access control to the corporate resources. Thus, the first step to implement ZTA is to identify what resources and assets the company owns, including corporate-owned and possibly BYOD devices. Making device inventory and user database helps the company keep track of this information. Then, the company needs to implement two mechanisms to identify devices requesting resources and to identify the user sending the request.

6.2. Removing Implicit Trust. Next, we remove implicit trust from all related subjects. All devices located both inside and outside the corporate network are treated the same as devices connected from the outside (external network). An enterprise may also utilize segmentation techniques, such as VLAN, to temporarily separate devices into safe and quarantine VLAN for better management.

6.3. Externalizing Workflows. In this step, candidate applications and workflows are externalized via internet-facing policy enforcing point (PEP) (see Figure 1). All requests passing the authentication and authorization processes are delegated to the backend server. In this step, policy engine (PE) performs a service-level authorization to grant or deny access to corporate resources. All communications in this step are encrypted.

6.4. A Case Study of Google's BeyondCorp. An excellent example of ZTA implementation is Google's BeyondCorp. In this section, we discuss how Google achieved its idea and implementation of the zero trust model according to steps explained earlier in Sections 6.1–6.3. Figure 3 shows the overview of all components in Google's BeyondCorp.

First, in BeyondCorp, Google allows only managed devices to communicate and send requests for corporate resources. A device inventory database responsible for storing device-related information was created to keep track of all managed device states and all relevant information. Also, digital certificates and some encryption keys are stored in TPM or qualified applications to help in securing and identifying these devices.

To identify a user, a multifactor authentication (MFA) is required. MFA can be performed using a managed device either via a single sign-on (SSO) platform against the predetermined user/group database or via RADIUS using 802.1x protocol.

Next, to remove implicit trust from the legacy network, some managed devices located in the private privilege network inside the Google building are then moved to an unprivileged network. These devices are now treated the same as devices connected from the external (outside) network. Utilizing VLAN, managed devices authenticated via RADIUS server using 802.1x protocol are assigned to a different virtual network. On the other hand, all unrecognized and unmanaged devices are automatically assigned to a guest or quarantine network for further remediation.

To externalize a workflow making it accessible from anywhere, Google deploys an internet-facing access proxy which is similar to policy enforcement point (PEP) in NIST SP 800–207 [2]. Applications in BeyondCorp are registered with public DNS having their CNAME pointing to Google's access proxy server. Hence, all requests coming from both public (external) and private networks are sent to the access proxy. After passing the authentication, the access proxy then asks the access control engine to verify and authorize/deny the request.

The access control engine (ACE, for short), equivalent to policy enforcement (PE) in [2], is considered the heart and the brain of Google's BeyondCorp. ACE incorporates various sources of information to decide if the requesting user and device are trustworthy and have the right to access the resources being requested or not. Every decision is made on a per-request basis.

Trust inference, similar to the trust algorithm in [2], is a logical component that continuously computes and updates the trust level of all subjects in real-time. An access request is authorized by the ACE only when all predefined rules/conditions are met, and the trust level of the requesting user is higher than the minimum required trust level defined for the resource being requested.

Regarding migration to ZTA, Google first identifies candidate workflow by performing workflow qualification, job function analysis, identifying candidate population, and traffic analysis via both privileged network traffic sampling and unprivileged network simulation. The migration starts with low-risk migration and moves to more critical workflows with higher risk once the company accumulates enough confidence in their migration process. More details regarding Google's BeyondCorp are provided in [19, 22–25].

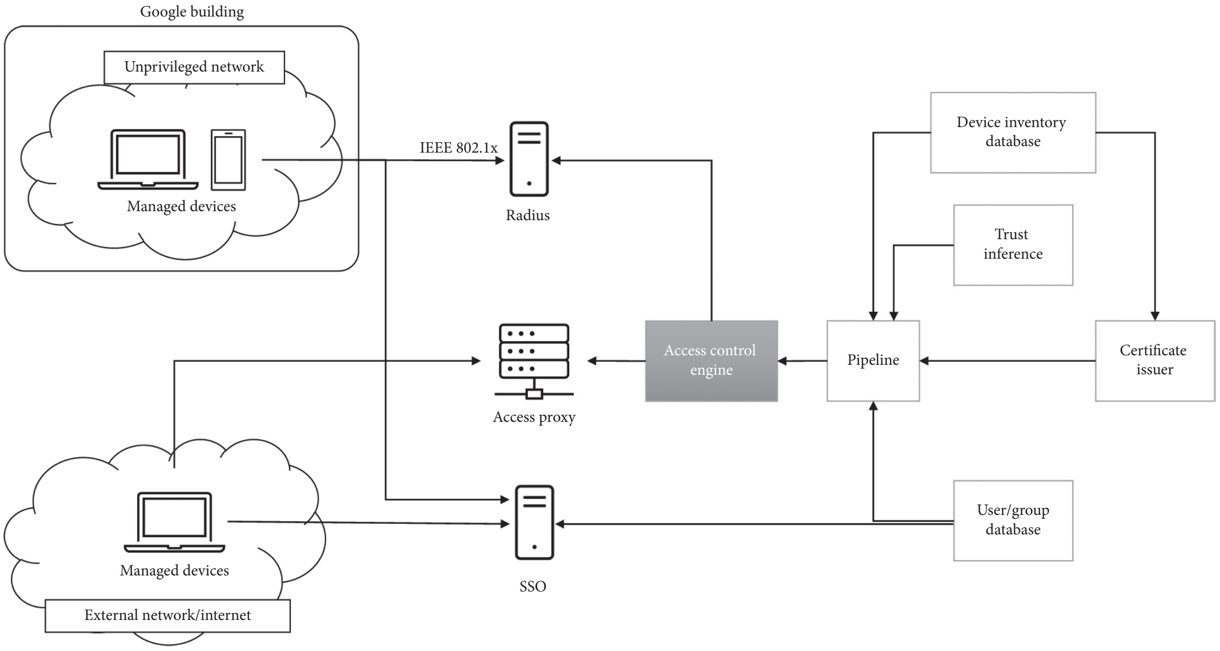

FIGURE 3: An overview of Google's BeyondCorp (original image from [22]).

7. Conclusions

In this paper, we discuss the concept and application of the zero trust architecture. Some challenges in ZTA, including lacking standardization and vendor lock-in problems, are introduced and discussed. Lastly, brief information focusing on steps and things to consider regarding migration from perimeter-based architecture to ZTA is presented.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] FireEye, "Highly evasive attacker leverages solarwinds supply chain to compromise multiple global victims with sunburst backdoor," 2020, https://www.fireeye.com/blog/threat-research/2020/12/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor.html.

[2] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, *Special Publication 800-207: Zero Trust Architecture*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2020, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf.

[3] Square Enix, "Square enix to make work from home permanent as of December 1 -mostly home-based hybrid model to strike balance between flexibility and manageability," 2020, https://www.jp.square-enix.com/company/en/news/2020/html/df9995782da2d516db9ebac425d02d4019665f70.html.

[4] C. Page, "Square enix expects 80% of employees to work from home permanently," 2020, https://www.forbes.com/sites/carlypage/2020/11/25/square-enix-expects-80-of-employees-to-work-from-home-permanently/?sh=60d3ed42294c.

[5] B. Chen, S. Qiao, J. Zhao et al., "A security awareness and protection system for 5G smart healthcare based on zero-trust architecture," *IEEE Internet of Things Journal*, 2020.

[6] Q. Yao, Q. Wang, X. Zhang, and J. Fei, "Dynamic access control and authorization system based on zero-trust architecture," in *Proceedings of the 2020 International Conference on Control, Robotics and Intelligent System*, Xiamen, China, October 2020.

[7] R. Vanickis, P. Jacob, S. Dehghanzadeh, and B. Lee, "Access control policy enforcement for zero-trust-networking," in *Proceedings of the 2018 29th Irish Signals and Systems Conference (ISSC)*, Belfast, UK, June 2018.

[8] OASIS, *eXtensible Access Control Markup Language (XACML)*, OASIS, Burlington, MA, USA, 2013, http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html.

[9] B. Embrey, "The top three factors driving zero trust adoption," *Computer Fraud & Security*, vol. 2020, no. 9, pp. 13–15, 2020.

[10] T. Pandit, "Cloud desktops further the shift to zero trust," 2019, https://www.workspot.com/blog/cloud-desktops-zero-trust/.

[11] Information Security Media Group, "Zero trust and the role of internet isolation," Information Security Media Group, Boston, UK, 2013, https://www.bankinfosecurity.com/zero-trust-role-internet-isolation-a-15652.

[12] I. Brosso, A. L. Neve, G. Bressan, and W. V. Ruggiero, "A continuous authentication system based on user behavior analysis," in *Proceedings of the 2010 International Conference on Availability, Reliability and Security*, Krakow, Poland, February 2010.

[13] Joint Task Force, *Special Publication 800-37 Rev. 2: Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2018, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf.

[14] Cybersecurity & Infrastructure Security Agency, "Identity, credential, and access management (ICAM)," Cybersecurity & Infrastructure Security Agency, Arlington, VA, USA, 2021, https://www.cisa.gov/safecom/icam-resources.

[15] P. A. Grassi, M. E. Garcia, and J. L. Fenton, *Special Publication 800-63-3: Digital Identity Guidelines*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2017, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf.

[16] U.S. Department of Health & Human Services, *Health Information Privacy*, U.S. Department of Health & Human Services, Washington, DC, USA, 2021, https://www.hhs.gov/hipaa/index.html.

[17] L. Goodspeed, *Request for Comments: PCI DSS Version 4.0 Draft Standard*, PCI Security Standards Council, Wakefield, MA, USA, 2020, https://blog.pcisecuritystandards.org/request-for-comments-pci-dss-version-4.0-draft-standard.

[18] The European Parliament and The Council of the European Union, *Regulation (EU) 2016/679 OF The European Parliament and of The Council of 27 April 2016 on the protection of Natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, European Union, Brussels, Belgium, 2016, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[19] J. Opara-Martins, R. Sahandi, and F. Tian, "Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective," *Journal of Cloud Computing*, vol. 5, no. 4, 2016.

[20] OASIS Cyber Threat Intelligence (CTI), *STIX™ version 2.1*, OASIS, Burlington, MA, USA, 2021, https://www.oasis-open.org/standard/6426/.

[21] OASIS, *TAXII™ Version 2.1*, OASIS, Burlington, MA, USA, 2021, https://docs.oasis-open.org/cti/taxii/v2.1/cs01/taxii-v2.1-cs01.pdf.

[22] R. Ward and B. Beyer, "BeyondCorp: a new approach to enterprise security," *Usenix*, vol. 39, no. 6, pp. 6–11, 2014, https://www.usenix.org/publications/login/dec14/ward.

[23] B. Osborn, J. McWilliams, B. Beyer, and M. Saltonstall, "BeyondCorp: design to deployment at Google," *Usenix*, vol. 41, pp. 28–34, 2016, https://www.usenix.org/publications/login/spring2016/osborn.

[24] L. Cittadini, B. Spear, B. Beyer, and M. Saltonstall, "BeyondCorp: the access proxy," *Usenix*, vol. 41, no. 4, pp. 28–33, 2016, https://www.usenix.org/publications/login/winter2016/cittadini.

[25] J. Peck, B. Beyer, C. Beske, and M. Saltonstall, "Migrating to BeyondCorp: maintaining productivity while improving security," *Usenix*, vol. 42, no. 2, pp. 49–55, 2017, https://www.usenix.org/publications/login/summer2017/peck.

[26] V. Escobedo, B. Beyer, M. Saltonstall, and F. Żyźniewski, "BeyondCorp: the user experience," *Usenix*, vol. 42, no. 3, pp. 38–43, 2017, https://www.usenix.org/publications/login/fall2017/escobedo.

[27] H. King, M. Janosko, B. Beyer, and M. Saltonstall, "BeyondCorp 6: building a healthy fleet," *Usenix*, vol. 43, no. 3, pp. 24–30, 2018, https://www.usenix.org/system/files/login/articles/login_fall18_05_king.pdf.