NIST Special Publication 800-207

Zero Trust Architecture

Scott Rose
Oliver Borchert
Stu Mitchell
Sean Connelly

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-207

C O M P U T E R S E C U R I T Y

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 800-207

Zero Trust Architecture

Scott Rose
Oliver Borchert
Advanced Network Technologies Division
Information Technology Laboratory

Stu Mitchell
Stu2Labs
Stafford, VA

Sean Connelly
Cybersecurity & Infrastructure Security Agency
Department of Homeland Security

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-207

August 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Under Secretary of Commerce for Standards and Technology

Authority

This publication has been developed by NIST in accordance with its statutory responsibilities under the Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 *et seq.*, Public Law (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including minimum requirements for federal information systems, but such standards and guidelines shall not apply to national security systems without the express approval of appropriate federal officials exercising policy authority over such systems. This guideline is consistent with the requirements of the Office of Management and Budget (OMB) Circular A-130.

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, Director of the OMB, or any other federal official. This publication may be used by nongovernmental organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would, however, be appreciated by NIST.

National Institute of Standards and Technology Special Publication 800-207
Natl. Inst. Stand. Technol. Spec. Publ. 800-207, 59 pages (August 2020)
CODEN: NSPUE2

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-207

Comments on this publication may be submitted to:

National Institute of Standards and Technology
Attn: Advanced Network Technologies Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8920) Gaithersburg, MD 20899-8920
Email: zerotrust-arch@nist.gov

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems. The Special Publication 800-series reports on ITL's research, guidelines, and outreach efforts in information system security, and its collaborative activities with industry, government, and academic organizations.

Abstract

Zero trust (ZT) is the term for an evolving set of cybersecurity paradigms that move defenses from static, network-based perimeters to focus on users, assets, and resources. A zero trust architecture (ZTA) uses zero trust principles to plan industrial and enterprise infrastructure and workflows. Zero trust assumes there is no implicit trust granted to assets or user accounts based solely on their physical or network location (i.e., local area networks versus the internet) or based on asset ownership (enterprise or personally owned). Authentication and authorization (both subject and device) are discrete functions performed before a session to an enterprise resource is established. Zero trust is a response to enterprise network trends that include remote users, bring your own device (BYOD), and cloud-based assets that are not located within an enterprise-owned network boundary. Zero trust focuses on protecting resources (assets, services, workflows, network accounts, etc.), not network segments, as the network location is no longer seen as the prime component to the security posture of the resource. This document contains an abstract definition of zero trust architecture (ZTA) and gives general deployment models and use cases where zero trust could improve an enterprise's overall information technology security posture.

Keywords

architecture; cybersecurity; enterprise; network security; zero trust.

Acknowledgments

This document is the product of a collaboration between multiple federal agencies and is overseen by the Federal CIO Council. The architecture subgroup is responsible for development of this document, but there are specific individuals who deserve recognition. These include Greg Holden, project manager of the Federal CIO Council ZTA project; Alper Kerman, project manager for the NIST/National Cybersecurity Center of Excellence ZTA effort; and Douglas Montgomery.

Audience

This document is intended to describe zero trust for enterprise security architects. It is meant to aid understanding of zero trust for civilian unclassified systems and provide a road map to migrate and deploy zero trust security concepts to an enterprise environment. Agency cybersecurity managers, network administrators, and managers may also gain insight into zero trust and ZTA from this document. It is not intended to be a single deployment plan for ZTA as an enterprise will have unique business use cases and data assets that require safeguards. Starting with a solid understanding of the organization's business and data will result in a strong approach to zero trust.

Trademark Information

Patent Disclosure Notice

NOTICE: The Information Technology Laboratory (ITL) has requested that holders of patent claims whose use may be required for compliance with the guidance or requirements of this publication disclose such patent claims to ITL. However, holders of patents are not obligated to respond to ITL calls for patents and ITL has not undertaken a patent search in order to identify which, if any, patents may apply to this publication.

Following the ITL call for the identification of patent claims whose use may be required for compliance with the guidance or requirements of this publication, notice of one or more such claims has been received.

By publication, no position is taken by ITL with respect to the validity or scope of any patent claim or of any rights in connection therewith. The known patent holder(s) has (have), however, provided to NIST a letter of assurance stating either (1) a general disclaimer to the effect that it does (they do) not hold and does (do) not currently intend holding any essential patent claim(s), or (2) that it (they) will negotiate royalty-free or royalty-bearing licenses with other parties on a demonstrably nondiscriminatory basis with reasonable terms and conditions.

Details may be obtained from **zerotrust-arch@nist.gov.**

No representation is made or implied that this is the only license that may be required to avoid patent infringement in the use of this publication.

Table of Contents

List of Appendices

List of Figures

List of Tables

1 Introduction

A typical enterprise's infrastructure has grown increasingly complex. A single enterprise may operate several internal networks, remote offices with their own local infrastructure, remote and/or mobile individuals, and cloud services. This complexity has outstripped legacy methods of perimeter-based network security as there is no single, easily identified perimeter for the enterprise. Perimeter-based network security has also been shown to be insufficient since once attackers breach the perimeter, further lateral movement is unhindered.

This complex enterprise has led to the development of a new model for cybersecurity known as "zero trust" (ZT). A ZT approach is primarily focused on data and service protection but can and should be expanded to include all enterprise assets (devices, infrastructure components, applications, virtual and cloud components) and subjects (end users, applications and other non-human entities that request information from resources). Throughout this document, "subject" will be used unless the section relates directly to a human end user in which "user" will be specifically used instead of the more generic "subject." Zero trust security models assume that an attacker is present in the environment and that an enterprise-owned environment is no different—or no more trustworthy—than any nonenterprise-owned environment. In this new paradigm, an enterprise must assume no implicit trust and continually analyze and evaluate the risks to its assets and business functions and then enact protections to mitigate these risks. In zero trust, these protections usually involve minimizing access to resources (such as data and compute resources and applications/services) to only those subjects and assets identified as needing access as well as continually authenticating and authorizing the identity and security posture of each access request.

A zero trust architecture (ZTA) is an enterprise cybersecurity architecture that is based on zero trust principles and designed to prevent data breaches and limit internal lateral movement. This publication discusses ZTA, its logical components, possible deployment scenarios, and threats. It also presents a general road map for organizations wishing to migrate to a zero trust design approach and discusses relevant federal policies that may impact or influence a zero trust architecture.

ZT is not a single architecture but a set of guiding principles for workflow, system design and operations that can be used to improve the security posture of any classification or sensitivity level [FIPS199]. Transitioning to ZTA is a journey concerning how an organization evaluates risk in its mission and cannot simply be accomplished with a wholesale replacement of technology. That said, many organizations already have elements of a ZTA in their enterprise infrastructure today. Organizations should seek to incrementally implement zero trust principles, process changes, and technology solutions that protect their data assets and business functions by use case. Most enterprise infrastructures will operate in a hybrid zero trust/perimeter-based mode while continuing to invest in IT modernization initiatives and improve organization business processes.

Organizations need to implement comprehensive information security and resiliency practices for zero trust to be effective. When balanced with existing cybersecurity policies and guidance, identity and access management, continuous monitoring, and best practices, a ZTA can protect

against common threats and improve an organization's security posture by using a managed risk approach.

1.1 History of Zero Trust Efforts Related to Federal Agencies

The concept of zero trust has been present in cybersecurity since before the term "zero trust" was coined. The Defense Information Systems Agency (DISA) and the Department of Defense published their work on a more secure enterprise strategy dubbed "black core" [BCORE]. Black core involved moving from a perimeter-based security model to one that focused on the security of individual transactions. The work of the Jericho Forum in 2004 publicized the idea of de-perimeterization—limiting implicit trust based on network location and the limitations of relying on single, static defenses over a large network segment [JERICHO]. The concepts of de-perimeterization evolved and improved into the larger concept of zero trust, which was later coined by John Kindervag[1] while at Forrester.[2] Zero trust then became the term used to describe various cybersecurity solutions that moved security away from implied trust based on network location and instead focused on evaluating trust on a per-transaction basis. Both private industry and higher education have also undergone this evolution from perimeter-based security to a security strategy based on zero trust principles.

Federal agencies have been urged to move to security based on zero trust principles for more than a decade, building capabilities and policies such as the Federal Information Security Modernization Act (FISMA) followed by the Risk Management Framework (RMF); Federal Identity, Credential, and Access Management (FICAM); Trusted Internet Connections (TIC); and Continuous Diagnostics and Mitigation (CDM) programs. All of these programs aim to restrict data and resource access to authorized parties. When these programs were started, they were limited by the technical capabilities of information systems. Security policies were largely static and were enforced at large "choke points" that an enterprise could control to get the largest effect for the effort. As technology matures, it is becoming possible to continually analyze and evaluate access requests in a dynamic and granular fashion to a "need to access" basis to mitigate data exposure due to compromised accounts, attackers monitoring a network, and other threats.

1.2 Structure of This Document

The rest of the document is organized as follows:

- **Section 2** defines ZT and ZTA and lists some assumptions when designing a ZTA for an enterprise. This section also includes a list of the tenets of ZT design.
- **Section 3** documents the logical components, or building blocks, of ZT. It is possible that unique implementations compose ZTA components differently yet serve the same logical functionality.

[1] https://go.forrester.com/blogs/next-generation-access-and-zero-trust/
[2] Any mention of commercial products or services within NIST documents is for information only; it does not imply a recommendation or endorsement by NIST.

- **Section 4** lists some possible use cases where a ZTA may make enterprise environments more secure and less prone to successful exploitation. These include enterprises with remote employees, cloud services, and guest networks.
- **Section 5** discusses threats to an enterprise using a ZTA. Many of these threats are similar to any architected networks but may require different mitigation techniques.
- **Section 6** discusses how ZTA tenets fit into and/or complement existing guidance for federal agencies.
- **Section 7** presents the starting point for transitioning an enterprise (such as a federal agency) to a ZTA. This includes a description of the general steps needed to plan and deploy applications and enterprise infrastructure that are guided by ZT tenets.

2 Zero Trust Basics

Zero trust is a cybersecurity paradigm focused on resource protection and the premise that trust is never granted implicitly but must be continually evaluated. Zero trust architecture is an end-to-end approach to enterprise resource and data security that encompasses identity (person and non-person entities), credentials, access management, operations, endpoints, hosting environments, and the interconnecting infrastructure. The initial focus should be on restricting resources to those with a need to access and grant only the minimum privileges (e.g., read, write, delete) needed to perform the mission. Traditionally, agencies (and enterprise networks in general) have focused on perimeter defense and authenticated subjects are given authorized access to a broad collection of resources once on the internal network. As a result, unauthorized lateral movement within the environment has been one of the biggest challenges for federal agencies.

The Trusted Internet Connections (TIC) and agency perimeter firewalls provide strong internet gateways. This helps block attackers from the internet, but the TICs and perimeter firewalls are less useful for detecting and blocking attacks from inside the network and cannot protect subjects outside of the enterprise perimeter (e.g., remote workers, cloud-based services, edge devices, etc.).

An operative definition of zero trust and zero trust architecture is as follows:

> *Zero trust* (ZT) provides a collection of concepts and ideas designed to minimize uncertainty in enforcing accurate, least privilege per-request access decisions in information systems and services in the face of a network viewed as compromised. *Zero trust architecture* (ZTA) is an enterprise's cybersecurity plan that utilizes zero trust concepts and encompasses component relationships, workflow planning, and access policies. Therefore, a zero trust enterprise is the network infrastructure (physical and virtual) and operational policies that are in place for an enterprise as a product of a zero trust architecture plan.

An enterprise decides to adopt zero trust as its core strategy and generate a zero trust architecture as a plan developed with zero trust principles (see Section 2.1 below) in mind. This plan is then deployed to produce a zero trust environment for use in the enterprise.

This definition focuses on the crux of the issue, which is the goal to *prevent unauthorized access to data and services* coupled with making the *access control enforcement as granular as possible*. That is, authorized and approved subjects (combination of user, application (or service), and device) can access the data to the exclusion of all other subjects (i.e., attackers). To take this one step further, the word "resource" can be substituted for "data" so that ZT and ZTA are about resource access (e.g., printers, compute resources, Internet of Things [IoT] actuators) and not just data access.

To lessen uncertainties (as they cannot be eliminated), the focus is on authentication, authorization, and shrinking implicit trust zones while maintaining availability and minimizing temporal delays in authentication mechanisms. Access rules are made as granular as possible to enforce least privileges needed to perform the action in the request.

In the abstract model of access shown in Figure 1, a subject needs access to an enterprise resource. Access is granted through a policy decision point (PDP) and corresponding policy enforcement point (PEP).[3]

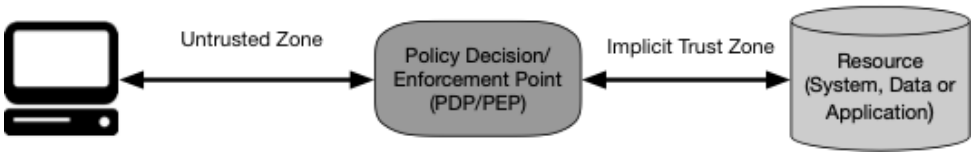

Figure 1: Zero Trust Access

The system must ensure that the subject is authentic and the request is valid. The PDP/PEP passes proper judgment to allow the subject to access the resource. This implies that zero trust applies to two basic areas: authentication and authorization. What is the level of confidence about the subject's identity for this unique request? Is access to the resource allowable given the level of confidence in the subject's identity? Does the device used for the request have the proper security posture? Are there other factors that should be considered and that change the confidence level (e.g., time, location of subject, subject's security posture)? Overall, enterprises need to develop and maintain dynamic risk-based policies for resource access and set up a system to ensure that these policies are enforced correctly and consistently for individual resource access requests. This means that an enterprise should not rely on implied trustworthiness wherein if the subject has met a base authentication level (e.g., logging into an asset), all subsequent resource requests are assumed to be equally valid.

The "implicit trust zone" represents an area where all the entities are trusted to at least the level of the last PDP/PEP gateway. For example, consider the passenger screening model in an airport. All passengers pass through the airport security checkpoint (PDP/PEP) to access the boarding gates. The passengers, airport employees, aircraft crew, etc., mill about in the terminal area, and all the individuals are considered trusted. In this model, the implicit trust zone is the boarding area.

The PDP/PEP applies a set of controls so that all traffic beyond the PEP has a common level of trust. The PDP/PEP cannot apply additional policies beyond its location in the flow of traffic. To allow the PDP/PEP to be as specific as possible, the implicit trust zone must be as small as possible.

Zero trust provides a set of principles and concepts around moving the PDP/PEPs closer to the resource. The idea is to explicitly authenticate and authorize all subjects, assets and workflows that make up the enterprise.

[3] Part of the concepts defined in OASIS XACML 2.0 https://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf

2.1 Tenets of Zero Trust

Many definitions and discussions of ZT stress the concept of removing wide-area perimeter defenses (e.g., enterprise firewalls) as a factor. However, most of these definitions continue to define themselves in relation to perimeters in some way (such as micro-segmentation or micro-perimeters; see Section 3.1) as part of the functional capabilities of a ZTA. The following is an attempt to define ZT and ZTA in terms of basic tenets that should be involved rather than what is excluded. These tenets are the ideal goal, though it must be acknowledged that not all tenets may be fully implemented in their purest form for a given strategy.

A zero trust architecture is designed and deployed with adherence to the following zero trust basic tenets:

1. **All data sources and computing services are considered resources.** A network may be composed of multiple classes of devices. A network may also have small footprint devices that send data to aggregators/storage, software as a service (SaaS), systems sending instructions to actuators, and other functions. Also, an enterprise may decide to classify personally owned devices as resources if they can access enterprise-owned resources.
2. **All communication is secured regardless of network location.** Network location alone does not imply trust. Access requests from assets located on enterprise-owned network infrastructure (e.g., inside a legacy network perimeter) must meet the same security requirements as access requests and communication from any other nonenterprise-owned network. In other words, trust should not be automatically granted based on the device being on enterprise network infrastructure. All communication should be done in the most secure manner available, protect confidentiality and integrity, and provide source authentication.
3. **Access to individual enterprise resources is granted on a per-session basis.** Trust in the requester is evaluated before the access is granted. Access should also be granted with the least privileges needed to complete the task. This could mean only "sometime recently" for this particular transaction and may not occur directly before initiating a session or performing a transaction with a resource. However, authentication and authorization to one resource will not automatically grant access to a different resource.
4. **Access to resources is determined by dynamic policy—including the observable state of client identity, application/service, and the requesting asset—and may include other behavioral and environmental attributes.** An organization protects resources by defining what resources it has, who its members are (or ability to authenticate users from a federated community), and what access to resources those members need. For zero trust, client identity can include the user account (or service identity) and any associated attributes assigned by the enterprise to that account or artifacts to authenticate automated tasks. Requesting asset state can include device characteristics such as software versions installed, network location, time/date of request, previously observed behavior, and installed credentials. Behavioral attributes include, but not limited to, automated subject analytics, device analytics, and measured deviations from observed usage patterns. Policy is the set of access rules based on attributes that an organization assigns to a subject, data asset, or application. Environmental attributes may include such factors as requestor

network location, time, reported active attacks, etc. These rules and attributes are based on the needs of the business process and acceptable level of risk. Resource access and action permission policies can vary based on the sensitivity of the resource/data. Least privilege principles are applied to restrict both visibility and accessibility.

5. **The enterprise monitors and measures the integrity and security posture of all owned and associated assets.** No asset is inherently trusted. The enterprise evaluates the security posture of the asset when evaluating a resource request. An enterprise implementing a ZTA should establish a continuous diagnostics and mitigation (CDM) or similar system to monitor the state of devices and applications and should apply patches/fixes as needed. Assets that are discovered to be subverted, have known vulnerabilities, and/or are not managed by the enterprise may be treated differently (including denial of all connections to enterprise resources) than devices owned by or associated with the enterprise that are deemed to be in their most secure state. This may also apply to associated devices (e.g., personally owned devices) that may be allowed to access some resources but not others. This, too, requires a robust monitoring and reporting system in place to provide actionable data about the current state of enterprise resources.
6. **All resource authentication and authorization are dynamic and strictly enforced before access is allowed.** This is a constant cycle of obtaining access, scanning and assessing threats, adapting, and continually reevaluating trust in ongoing communication. An enterprise implementing a ZTA would be expected to have Identity, Credential, and Access Management (ICAM) and asset management systems in place. This includes the use of multifactor authentication (MFA) for access to some or all enterprise resources. Continual monitoring with possible reauthentication and reauthorization occurs throughout user transactions, as defined and enforced by policy (e.g., time-based, new resource requested, resource modification, anomalous subject activity detected) that strives to achieve a balance of security, availability, usability, and cost-efficiency.
7. **The enterprise collects as much information as possible about the current state of assets, network infrastructure and communications and uses it to improve its security posture.** An enterprise should collect data about asset security posture, network traffic and access requests, process that data, and use any insight gained to improve policy creation and enforcement. This data can also be used to provide context for access requests from subjects (see Section 3.3.1).

The above tenets attempt to be technology agnostic. For example, “user identification (ID)” could include several factors such as username/password, certificates, and onetime password. These tenets apply to work done within an organization or in collaboration with one or more partner organizations and not to anonymous public or consumer-facing business processes. An organization cannot impose internal policies on external actors (e.g., customers or general internet users) but may be able to implement some ZT-based policies on nonenterprise users who have a special relationship with the organization (e.g. registered customers, employee dependents, etc.).

2.2 A Zero Trust View of a Network

There are some basic assumptions for network connectivity for any organization that utilizes ZTA in network planning and deployment. Some of these assumptions apply to enterprise-owned network infrastructure, and some apply to enterprise-owned resources operating on nonenterprise-owned network infrastructure (e.g., public Wi-Fi or public cloud providers). These assumptions are used to direct the formation of a ZTA. The network in an enterprise implementing a ZTA should be developed with the ZTA tenets outlined above and with the following assumptions.

1. **The entire enterprise private network is not considered an implicit trust zone.** Assets should always act as if an attacker is present on the enterprise network, and communication should be done in the most secure manner available (see tenet 2 above). This entails actions such as authenticating all connections and encrypting all traffic.
2. **Devices on the network may not be owned or configurable by the enterprise.** Visitors and/or contracted services may include nonenterprise-owned assets that need network access to perform their role. This includes bring-your-own-device (BYOD) policies that allow enterprise subjects to use nonenterprise-owned devices to access enterprise resources.
3. **No resource is inherently trusted.** Every asset must have its security posture evaluated via a PEP before a request is granted to an enterprise-owned resource (similar to tenet 6 above for assets as well as subjects). This evaluation should be continual for as long as the session lasts. Enterprise-owned devices may have artifacts that enable authentication and provide a confidence level higher than the same request coming from nonenterprise-owned devices. Subject credentials alone are insufficient for device authentication to an enterprise resource.
4. **Not all enterprise resources are on enterprise-owned infrastructure.** Resources include remote enterprise subjects as well as cloud services. Enterprise-owned or -managed assets may need to utilize the local (i.e., nonenterprise) network for basic connectivity and network services (e.g., DNS resolution).
5. **Remote enterprise subjects and assets cannot fully trust their local network connection.** Remote subjects should assume that the local (i.e., nonenterprise-owned) network is hostile. Assets should assume that all traffic is being monitored and potentially modified. All connection requests should be authenticated and authorized, and all communications should be done in the most secure manner possible (i.e., provide confidentiality, integrity protection, and source authentication). See the tenets of ZTA above.
6. **Assets and workflows moving between enterprise and nonenterprise infrastructure should have a consistent security policy and posture.** Assets and workloads should retain their security posture when moving to or from enterprise-owned infrastructure. This includes devices that move from enterprise networks to nonenterprise networks (i.e. remote users). This also includes workloads migrating from on-premises data centers to nonenterprise cloud instances.

3 Logical Components of Zero Trust Architecture

There are numerous logical components that make up a ZTA deployment in an enterprise. These components may be operated as an on-premises service or through a cloud-based service. The conceptual framework model in Figure 2 shows the basic relationship between the components and their interactions. Note that this is an ideal model showing logical components and their interactions. From Figure 1, the policy decision point (PDP) is broken down into two logical components: the policy engine and policy administrator (defined below). The ZTA logical components use a separate control plane to communicate, while application data is communicated on a data plane (see Section 3.4).

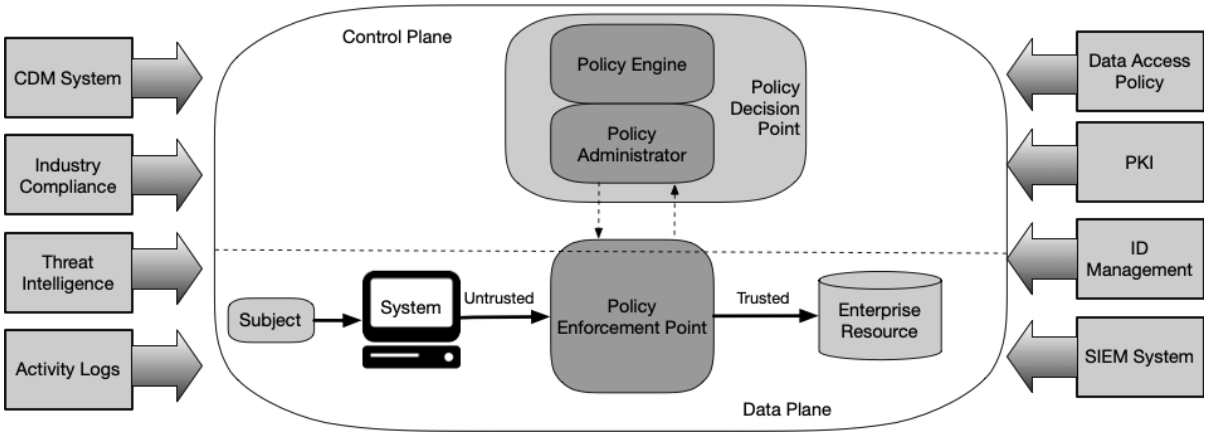

Figure 2: Core Zero Trust Logical Components

The component descriptions:

- **Policy engine (PE):** This component is responsible for the ultimate decision to grant access to a resource for a given subject. The PE uses enterprise policy as well as input from external sources (e.g., CDM systems, threat intelligence services described below) as input to a trust algorithm (see Section 3.3 for more details) to grant, deny, or revoke access to the resource. The PE is paired with the policy administrator component. The policy engine makes and logs the decision (as approved, or denied), and the policy administrator executes the decision.
- **Policy administrator (PA):** This component is responsible for establishing and/or shutting down the communication path between a subject and a resource (via commands to relevant PEPs). It would generate any session-specific authentication and authentication token or credential used by a client to access an enterprise resource. It is closely tied to the PE and relies on its decision to ultimately allow or deny a session. If the session is authorized and the request authenticated, the PA configures the PEP to allow the session to start. If the session is denied (or a previous approval is countermanded), the PA signals to the PEP to shut down the connection. Some implementations may treat the PE and PA as a single service; here, it is divided into its

two logical components. The PA communicates with the PEP when creating the communication path. This communication is done via the control plane.

- **Policy enforcement point (PEP):** This system is responsible for enabling, monitoring, and eventually terminating connections between a subject and an enterprise resource. The PEP communicates with the PA to forward requests and/or receive policy updates from the PA. This is a single logical component in ZTA but may be broken into two different components: the client (e.g., agent on a laptop) and resource side (e.g., gateway component in front of resource that controls access) or a single portal component that acts as a gatekeeper for communication paths. Beyond the PEP is the trust zone (see Section 2) hosting the enterprise resource.

In addition to the core components in an enterprise implementing a ZTA, several data sources provide input and policy rules used by the policy engine when making access decisions. These include local data sources as well as external (i.e., nonenterprise-controlled or -created) data sources. These can include:

- **Continuous diagnostics and mitigation (CDM) system:** This gathers information about the enterprise asset's current state and applies updates to configuration and software components. An enterprise CDM system provides the policy engine with the information about the asset making an access request, such as whether it is running the appropriate patched operating system (OS), the integrity of enterprise-approved software components or presence of non-approved components and whether the asset has any known vulnerabilities. CDM systems are also responsible for identifying and potentially enforcing a subset of polices on nonenterprise devices active on enterprise infrastructure.
- **Industry compliance system:** This ensures that the enterprise remains compliant with any regulatory regime that it may fall under (e.g., FISMA, healthcare or financial industry information security requirements). This includes all the policy rules that an enterprise develops to ensure compliance.
- **Threat intelligence feed(s):** This provides information from internal or external sources that help the policy engine make access decisions. These could be multiple services that take data from internal and/or multiple external sources and provide information about newly discovered attacks or vulnerabilities. This also includes newly discovered flaws in software, newly identified malware, and reported attacks to other assets that the policy engine will want to deny access to from enterprise assets.
- **Network and system activity logs:** This enterprise system aggregates asset logs, network traffic, resource access actions, and other events that provide real-time (or near-real-time) feedback on the security posture of enterprise information systems.
- **Data access policies:** These are the attributes, rules, and policies about access to enterprise resources. This set of rules could be encoded in (via management interface) or dynamically generated by the policy engine. These policies are the starting point for authorizing access to a resource as they provide the basic access privileges for accounts and applications/services in the enterprise. These policies should be based on the defined mission roles and needs of the organization.

- **Enterprise public key infrastructure (PKI):** This system is responsible for generating and logging certificates issued by the enterprise to resources, subjects, services and applications. This also includes the global certificate authority ecosystem and the Federal PKI,[4] which may or may not be integrated with the enterprise PKI. This could also be a PKI that is not built upon X.509 certificates.
- **ID management system:** This is responsible for creating, storing, and managing enterprise user accounts and identity records (e.g., lightweight directory access protocol (LDAP) server). This system contains the necessary subject information (e.g., name, email address, certificates) and other enterprise characteristics such as role, access attributes, and assigned assets. This system often utilizes other systems (such as a PKI) for artifacts associated with user accounts. This system may be part of a larger federated community and may include nonenterprise employees or links to nonenterprise assets for collaboration.
- **Security information and event management (SIEM) system:** This collects security-centric information for later analysis. This data is then used to refine policies and warn of possible attacks against enterprise assets.

3.1 Variations of Zero Trust Architecture Approaches

There are several ways that an enterprise can enact a ZTA for workflows. These approaches vary in the components used and in the main source of policy rules for an organization. Each approach implements all the tenets of ZT (see Section 2.1) but may use one or two (or one component) as the main driver of policies. A full ZT solution will include elements of all three approaches. The approaches include enhanced identity governance–driven, logical micro-segmentation, and network-based segmentation.

Certain approaches lend themselves to some use cases more than others. An organization looking to develop a ZTA for its enterprise may find that its chosen use case and existing policies point to one approach over others. That does not mean the other approaches would not work but rather that other approaches may be more difficult to implement and may require more fundamental changes to how the enterprise currently conducts business flows.

3.1.1 ZTA Using Enhanced Identity Governance

The enhanced identity governance approach to developing a ZTA uses the identity of actors as the key component of policy creation. If it were not for subjects requesting access to enterprise resources, there would be no need to create access polices. For this approach, enterprise resource access policies are based on identity and assigned attributes. The primary requirement for resource access is based on the access privileges granted to the given subject. Other factors such as device used, asset status, and environmental factors may alter the final confidence level calculation (and ultimate access authorization) or tailor the result in some way, such as granting only partial access to a given data source based on network location. Individual resources or PEP

[4] https://www.idmanagement.gov/topics/fpki/

components protecting the resource must have a way to forward requests to a policy engine service or authenticate the subject and approve the request before granting access.

Enhanced identity governance-based approaches for enterprises are often employed using an open network model or an enterprise network with visitor access or frequent nonenterprise devices on the network (such as with the use case in Section 4.3 below). Network access is initially granted to all assets but access to enterprise resources are restricted to identities with the appropriate access privileges. There is a downside in granting basic network connectivity as malicious actors could still attempt network reconnaissance and/or use the network to launch denial of service attacks either internally or against a third party. Enterprises still need to monitor and respond to such behavior before it impacts workflows.

The identity-driven approach works well with the resource portal model (see Section 3.2.3) since device identity and status provide secondary support data to access decisions. Other models work as well, depending on policies in place. Identity-driven approaches also work well for enterprises that use cloud-based applications/services that may not allow for enterprise-owned or -operated ZT security components to be used (such as many SaaS offerings). The enterprise can use the identity of requestors to form and enforce policy on these platforms.

3.1.2 ZTA Using Micro-Segmentation

An enterprise may choose to implement a ZTA based on placing individual or groups of resources on a unique network segment protected by a gateway security component. In this approach, the enterprise places infrastructure devices such as intelligent switches (or routers) or next generation firewalls (NGFWs) or special purpose gateway devices to act as PEPs protecting each resource or small group of related resources. Alternatively (or additionally), the enterprise may choose to implement host-based micro-segmentation using software agents (see Section 3.2.1) or firewalls on the endpoint asset(s), These gateway devices dynamically grant access to individual requests from a client, asset or service. Depending on the model, the gateway may be the sole PEP component or part of a multipart PEP consisting of the gateway and client-side agent (see Section 3.2.1).

This approach applies to a variety of use cases and deployment models as the protecting device acts as the PEP, with management of said devices acting as the PE/PA component. This approach requires an identity governance program (IGP) to fully function but relies on the gateway components to act as the PEP that shields resources from unauthorized access and/or discovery.

The key necessity to this approach is that the PEP components are managed and should be able to react and reconfigure as needed to respond to threats or change in the workflow. It is possible to implement some features of a micro-segmented enterprise by using less advanced gateway devices and even stateless firewalls, but the administration cost and difficulty to quickly adapt to changes make this a very poor choice.

3.1.3 ZTA Using Network Infrastructure and Software Defined Perimeters

The last approach uses the network infrastructure to implement a ZTA. The ZTA implementation could be achieved by using an overlay network (i.e., layer 7 but also could be set up lower of the

OSI network stack). These approaches are sometimes referred to as software defined perimeter (SDP) approaches and frequently include concepts from Software Defined Networks (SDN) [SDNBOOK] and intent-based networking (IBN) [IBNVN]. In this approach, the PA acts as the network controller that sets up and reconfigures the network based on the decisions made by the PE. The clients continue to request access via PEPs, which are managed by the PA component.

When the approach is implemented at the application network layer (i.e., layer 7), the most common deployment model is the agent/gateway (see Section 3.2.1). In this implementation, the agent and resource gateway (acting as the single PEP and configured by the PA) establish a secure channel used for communication between the client and resource. There may be other variations of this model, as well for cloud virtual networks, non-IP based networks, etc.

3.2 Deployed Variations of the Abstract Architecture

All of the above components are logical components. They do not necessarily need to be unique systems. A single asset may perform the duties of multiple logical components, and likewise, a logical component may consist of multiple hardware or software elements to perform the tasks. For example, an enterprise-managed PKI may consist of one component responsible for issuing certificates for devices and another used for issuing certificates to end users, but both use intermediate certificates issued from the same enterprise root certificate authority. In some ZT product offerings currently available on the market, the PE and PA components are combined in a single service.

There are several variations on deployment of selected components of the architecture that are outlined in the sections below. Depending on how an enterprise network is set up, multiple ZTA deployment models may be in use for different business processes in one enterprise.

3.2.1 Device Agent/Gateway-Based Deployment

In this deployment model, the PEP is divided into two components that reside on the resource or as a component directly in front of a resource. For example, each enterprise-issued asset has an installed device agent that coordinates connections, and each resource has a component (i.e., gateway) that is placed directly in front so that the resource communicates only with the gateway, essentially serving as a proxy for the resource. The agent is a software component that directs some (or all) traffic to the appropriate PEP in order for requests to be evaluated. The gateway is responsible for communicating with the policy administrator and allowing only approved communication paths configured by the policy administrator (see Figure 3).

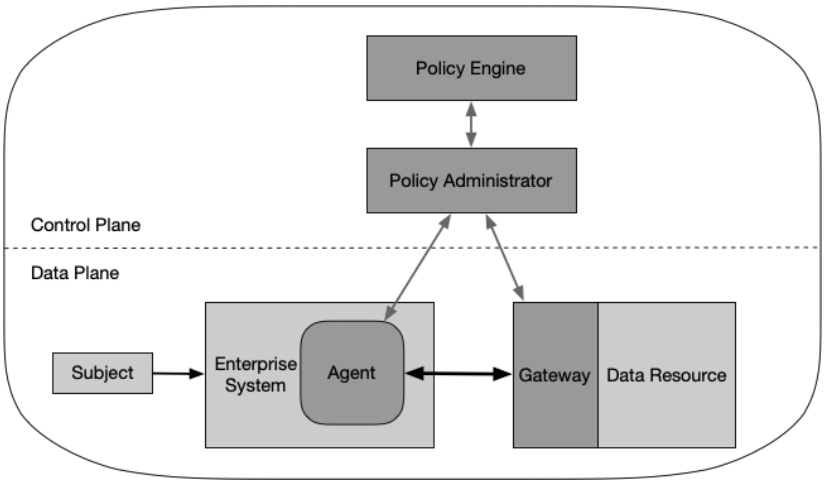

Figure 3: Device Agent/Gateway Model

In a typical scenario, a subject with an enterprise-issued laptop wishes to connect to an enterprise resource (e.g., human resources application/database). The access request is taken by the local agent, and the request is forwarded to the policy administrator. The policy administrator and policy engine could be an enterprise local asset or a cloud-hosted service. The policy administrator forwards the request to the policy engine for evaluation. If the request is authorized, the policy administrator configures a communication channel between the device agent and the relevant resource gateway via the control plane. This may include information such as an internet protocol (IP) address, port information, session key, or similar security artifacts. The device agent and gateway then connect, and encrypted application/service data flows begin. The connection between the device agent and resource gateway is terminated when the workflow is completed or when triggered by the policy administrator due to a security event (e.g., session time-out, failure to reauthenticate).

This model is best utilized for enterprises that have a robust device management program in place as well as discrete resources that can communicate with the gateway. For enterprises that heavily utilize cloud services, this is a client-server implementation of the Cloud Security Alliance (CSA) Software Defined Perimeter (SDP) [CSA-SDP]. This model is also appropriate for enterprises that do not want a BYOD policy in place. Access is possible only via the device agent, which can be placed on enterprise-owned assets.

3.2.2 Enclave-Based Deployment

This deployment model is a variation of the device agent/gateway model above. In this model, the gateway components may not reside on assets or in front of individual resources but instead reside at the boundary of a resource enclave (e.g., on-location data center) as shown in Figure 4. Usually, these resources serve a single business function or may not be able to communicate directly to a gateway (e.g., legacy database system that does not have an application programming interface [API] that can be used to communicate with a gateway). This deployment model may also be useful for enterprises that use cloud-based micro-services for a single business processes (e.g., user notification, database lookup, salary disbursement). In this model, the entire private cloud is located behind a gateway.

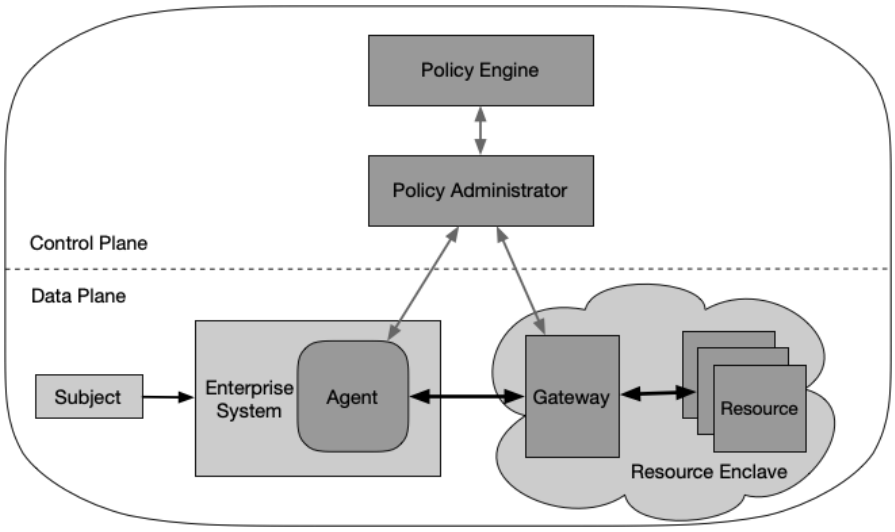

Figure 4: Enclave Gateway Model

It is possible for this model to be a hybrid with the device agent/gateway model. In this model, enterprise assets have a device agent that is used to connect to enclave gateways, but these connections are created using the same process as the basic device agent/gateway model.

This model is useful for enterprises that have legacy applications or on-premises data centers that cannot have individual gateways in place. The enterprise needs a robust asset and configuration management program in place to install/configure the device agents. The downside is that the gateway protects a collection of resources and may not be able to protect each resource individually. This may also allow for subjects to see resources which they do not have privileges to access.

3.2.3 Resource Portal-Based Deployment

In this deployment model, the PEP is a single component that acts as a gateway for subject requests. The gateway portal can be for an individual resource or a secure enclave for a collection of resources used for a single business function. One example would be a gateway portal into a private cloud or data center containing legacy applications as shown in Figure 5.

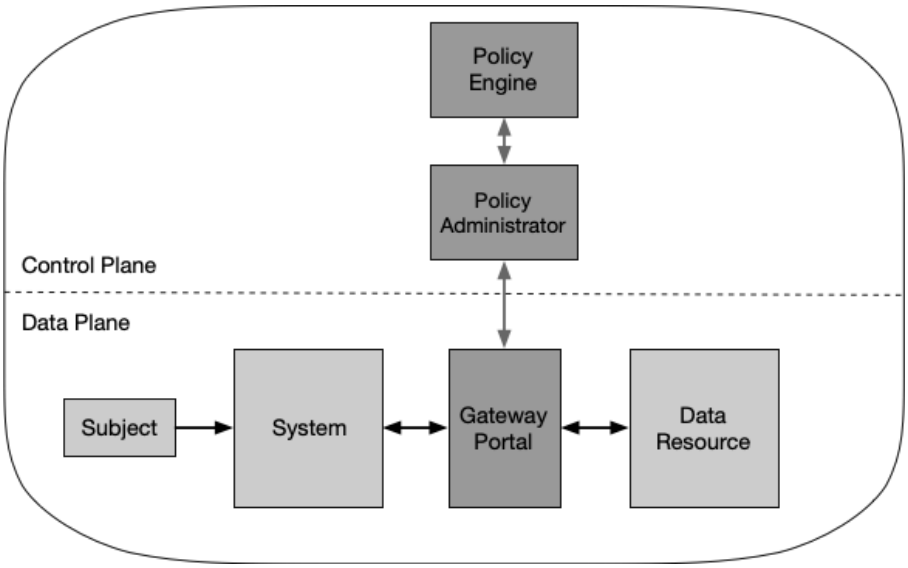

Figure 5: Resource Portal Model

The primary benefit of this model over the others is that a software component does not need to be installed on all client devices. This model is also more flexible for BYOD policies and inter-organizational collaboration projects. Enterprise administrators do not need to ensure that each device has the appropriate device agent before use. However, limited information can be inferred from devices requesting access. This model can only scan and analyze assets and devices once they connect to the PEP portal and may not be able to continuously monitor them for malware, unpatched vulnerabilities, and appropriate configuration.

The main difference with this model is there is no local agent that handles requests, and so the enterprise may not have full visibility or arbitrary control over assets as it can only see/scan them when they connect to a portal. The enterprise may be able to employ measures such as browser isolation to mitigate or compensate. These assets may be invisible to the enterprise between these sessions. This model also allows attackers to discover and attempt to access the portal or attempt a denial-of-service (DoS) attack against the portal. The portal systems should be well-provisioned to provide availability against a DoS attack or network disruption.

3.2.4 Device Application Sandboxing

Another variation of the agent/gateway deployment model is having vetted applications or processes run compartmentalized on assets. These compartments could be virtual machines, containers, or some other implementation, but the goal is the same: to protect the application or instances of applications from a possibly compromised host or other applications running on the asset.

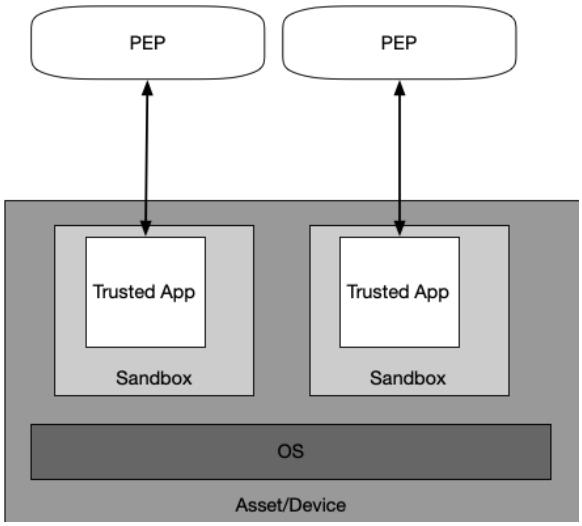

Figure 6: Application Sandboxes

In Figure 6, the subject device runs approved, vetted applications in a sandbox. The applications can communicate with the PEP to request access to resources, but the PEP will refuse requests from other applications on the asset. The PEP could be an enterprise local service or a cloud service in this model.

The main advantage of this model variant is that individual applications are segmented from the rest of the asset. If the asset cannot be scanned for vulnerabilities, these individual, sandboxed applications may be protected from a potential malware infection on the host asset. One of the disadvantages of this model is that enterprises must maintain these sandboxed applications for all assets and may not have full visibility into client assets. The enterprise also needs to make sure each sandboxed application is secure, which may require more effort than simply monitoring devices.

3.3 Trust Algorithm

For an enterprise with a ZTA deployment, the policy engine can be thought of as the brain and the PE's trust algorithm as its primary thought process. The trust algorithm (TA) is the process used by the policy engine to ultimately grant or deny access to a resource. The policy engine takes input from multiple sources (see Section 3): the policy database with observable information about subjects, subject attributes and roles, historical subject behavior patterns, threat intelligence sources, and other metadata sources. The process can be grouped into broad categories and visualized in Figure 7.

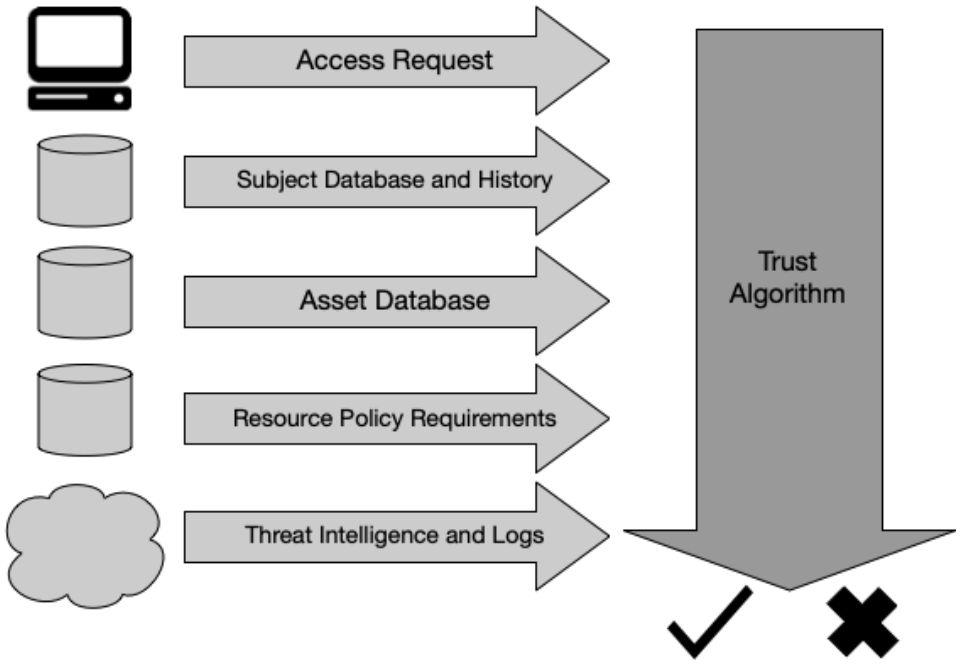

Figure 7: Trust Algorithm Input

In the figure, the inputs can be broken into categories based on what they provide to the trust algorithm.

- **Access request:** This is the actual request from the subject. The resource requested is the primary information used, but information about the requester is also used. This can include OS version, software used (e.g., does the requesting application appear on a list of approved applications?), and patch level. Depending on these factors and the asset security posture, access to assets might be restricted or denied.
- **Subject database:** This is the "who" that is requesting access to a resource [SP800-63]. This is the set of subjects (human and processes) of the enterprise or collaborators and a collection of subject attributes/privileges assigned. These subjects and attributes form the basis of policies for resource access [SP800-162] [NISTIR 7987]. User identities can include a mix of logical identity (e.g., account ID) and results of authentication checks performed by PEPs. Attributes of identity that can be factored into deriving the confidence level include time and geolocation. A collection of privileges given to multiple subjects could be thought of as a role, but privileges should be assigned to a subject on an individual basis and not simply because they may fit into a particular role in the organization. This collection should be encoded and stored in an ID management system and policy database. This may also include data about past observed subject behavior in some (TA) variants (see Section 3.3.1).
- **Asset database (and observable status):** This is the database that contains the known status of each enterprise-owned (and possibly known nonenterprise/BYOD) asset (physical and virtual, to some extent). This is compared to the observable status of the asset making the request and can include OS version, software present, and its integrity,

location (network location and geolocation), and patch level. Depending on the asset state compared with this database, access to assets might be restricted or denied.

- **Resource requirements:** This set of policies complements the user ID and attributes database [SP800-63] and defines the minimal requirements for access to the resource. Requirements may include authenticator assurance levels, such as MFA network location (e.g., deny access from overseas IP addresses), data sensitivity, and requests for asset configuration. These requirements should be developed by both the data custodian (i.e., those responsible for the data) and those responsible for the business processes that utilize the data (i.e., those responsible for the mission).
- **Threat intelligence:** This is an information feed or feeds about general threats and active malware operating on the internet. This could also include specific information about communication seen from the device that may be suspect (such as queries for possible malware command and control nodes). These feeds can be external services or internal scans and discoveries and can include attack signatures and mitigations. This is the only component that will most likely be under the control of a service rather than the enterprise.

The weight of importance for each data source may be a proprietary algorithm or may be configured by the enterprise. These weight values can be used to reflect the importance of the data source to an enterprise.

The final determination is then passed to the PA for execution. The PA's job is to configure the necessary PEPs to enable authorized communication. Depending on how the ZTA is deployed, this may involve sending authentication results and connection configuration information to gateways and agents or resource portals. PAs may also place a hold or pause on a communication session to reauthenticate and reauthorize the connection in accordance with policy requirements. The PA is also responsible for issuing the command to terminate the connection based on policy (e.g., after a time-out, when the workflow has been completed, due to a security alert).

3.3.1 Trust Algorithm Variations

There are different ways to implement a TA. Different implementers may wish to weigh the above factors differently according to the factors' perceived importance. There are two other major characteristics that can be used to differentiate TAs. The first is how the factors are evaluated, whether as binary decisions or weighted parts of a whole "score" or confidence level. The second is how requests are evaluated in relation to other requests by the same subject, application/service, or device.

- **Criteria- versus score-based:** A criteria-based TA assumes a set of qualified attributes that must be met before access is granted to a resource or an action (e.g., read/write) is allowed. These criteria are configured by the enterprise and should be independently configured for every resource. Access is granted or an action applied to a resource only if all the criteria are met. A score-based TA computes a confidence level based on values for every data source and enterprise-configured weights. If the score is greater than the configured threshold value for the resource, access is granted, or the action is performed.

Otherwise, the request is denied, or access privileges are reduced (e.g., read access is granted but not write access for a file).

- **Singular versus contextual:** A singular TA treats each request individually and does not take the subject history into consideration when making its evaluation. This can allow faster evaluations, but there is a risk that an attack can go undetected if it stays within a subject's allowed role. A contextual TA takes the subject or network agent's recent history into consideration when evaluating access requests. This means the PE must maintain some state information on all subjects and applications but may be more likely to detect an attacker using subverted credentials to access information in a pattern that is atypical of what the PE sees for the given subject. This also means that the PE must be informed of user behavior by the PA (and PEPs) that subjects interact with when communicating. Analysis of subject behavior can be used to provide a model of acceptable use, and deviations from this behavior could trigger additional authentication checks or resource request denials.

The two factors are not always dependent on each other. It is possible to have a TA that assigns a confidence level to every subject and/or device and still considers every access request independently (i.e., singular). However, contextual, score-based TAs would provide the ability to offer more dynamic and granular access control, since the score provides a current confidence level for the requesting account and adapts to changing factors more quickly than static policies modified by human administrators.

Ideally, a ZTA trust algorithm should be contextual, but this may not always be possible with the infrastructure components available to the enterprise. A contextual TA can mitigate threats where an attacker stays close to a "normal" set of access requests for a compromised subject account or insider attack. It is important to balance security, usability, and cost-effectiveness when defining and implementing trust algorithms. Continually prompting a subject for reauthentication against behavior that is consistent with historical trends and norms for their mission function and role within the organization can lead to usability issues. For example, if an employee in the HR department of an agency normally accesses 20 to 30 employee records in a typical workday, a contextual TA may send an alert if the access requests suddenly exceed 100 records in a day. A contextual TA may also send an alert if someone is making access requests after normal business hours as this could be an attacker exfiltrating records by using a compromised HR account. These are examples where a contextual TA can detect an attack whereas a singular TA may fail to detect the new behavior. In another example, an accountant who typically accesses the financial system during normal business hours is now trying to access the system in the middle of the night from an unrecognizable location. A contextual TA may trigger an alert and require the subject to satisfy a more stringent confidence level or other criteria as outlined in NIST Special Publication 800-63A [SP800-63A].

Developing a set of criteria or weights/threshold values for each resource requires planning and testing. Enterprise administrators may encounter issues during the initial implementation of ZTA where access requests that should be approved are denied due to misconfiguration. This will result in an initial "tuning" phase of deployment. Criteria or scoring weights may need to be adjusted to ensure that the policies are enforced while still allowing the enterprise's business processes to function. How long this tuning phase lasts depends on the enterprise-defined metrics

for progress and tolerance for incorrect access denials/approvals for the resources used in the workflow.

3.4 Network/Environment Components

In a ZT environment, there should be a separation (logical or possibly physical) of the communication flows used to control and configure the network and application/service communication flows used to perform the actual work of the organization. This is often broken down to a *control plane* for network control communication and a *data plane* for application/service communication flows [Gilman].

The control plane is used by various infrastructure components (both enterprise-owned and from service providers) to maintain and configure assets; judge, grant, or deny access to resources; and perform any necessary operations to set up communication paths between resources. The data plane is used for actual communication between software components. This communication channel may not be possible before the path has been established via the control plane. For example, the control plane could be used by the PA and PEP to set up the communication path between the subject and the enterprise resource. The application/service workload would then use the data plane path that was established.

3.4.1 Network Requirements to Support ZTA

1. **Enterprise assets have basic network connectivity.** The local area network (LAN), enterprise controlled or not, provides basic routing and infrastructure (e.g., DNS). The remote enterprise asset may not necessarily use all infrastructure services.
2. **The enterprise must be able to distinguish between what assets are owned or managed by the enterprise and the devices' current security posture.** This is determined by enterprise-issued credentials and not using information that cannot be authenticated information (e.g., network MAC addresses that can be spoofed).
3. **The enterprise can observe all network traffic.** The enterprise records packets seen on the data plane, even if it is not be able to perform application layer inspection (i.e., OSI layer 7) on all packets. The enterprise filters out metadata about the connection (e.g., destination, time, device identity) to dynamically update policies and inform the PE as it evaluates access requests.
4. **Enterprise resources should not be reachable without accessing a PEP.** Enterprise resources do not accept arbitrary incoming connections from the internet. Resources accept custom-configured connections only after a client has been authenticated and authorized. These communication paths are set up by the PEP. Resources may not even be discoverable without accessing a PEP. This prevents attackers from identifying targets via scanning and/or launching DoS attacks against resources located behind PEPs. Note that not all resources should be hidden this way; some network infrastructure components (e.g., DNS servers) must be accessible.
5. **The data plane and control plane are logically separate.** The policy engine, policy administrator, and PEPs communicate on a network that is logically separate and not directly accessible by enterprise assets and resources. The data plane is used for

application/service data traffic. The policy engine, policy administrator, and PEPs use the control plane to communicate and manage communication paths between assets. The PEPs must be able to send and receive messages from both the data and control planes.

6. **Enterprise assets can reach the PEP component.** Enterprise subjects must be able to access the PEP component to gain access to resources. This could take the form of a web portal, network device, or software agent on the enterprise asset that enables the connection.
7. **The PEP is the only component that accesses the policy administrator as part of a business flow.** Each PEP operating on the enterprise network has a connection to the policy administrator to establish communication paths from clients to resources. All enterprise business process traffic passes through one or more PEPs.
8. **Remote enterprise assets should be able to access enterprise resources without needing to traverse enterprise network infrastructure first.** For example, a remote subject should not be required to use a link back to the enterprise network (i.e., virtual private network [VPN]) to access services utilized by the enterprise and hosted by a public cloud provider (e.g., email).
9. **The infrastructure used to support the ZTA access decision process should be made scalable to account for changes in process load.** The PE(s), PA(s), and PEPs used in a ZTA become the key components in any business process. Delay or inability to reach a PEP (or inability of the PEPs to reach the PA/PE) negatively impacts the ability to perform the workflow. An enterprise implementing a ZTA needs to provision the components for the expected workload or be able to rapidly scale the infrastructure to handle increased usage when needed.
10. **Enterprise assets may not be able to reach certain PEPs due to policy or observable factors.** For example, there may be a policy stating that mobile assets may not be able to reach certain resources if the requesting asset is located outside of the enterprise's home country. These factors could be based on location (geolocation or network location), device type, or other criteria.

4 Deployment Scenarios/Use Cases

Any enterprise environment can be designed with zero trust tenets in mind. Most organizations already have some elements of zero trust in their enterprise infrastructure or are on their way through implementation of information security and resiliency policies and best practices. Several deployment scenarios and use cases lend themselves readily to a zero trust architecture. For instance, ZTA has its roots in organizations that are geographically distributed and/or have a highly mobile workforce. That said, any organization can benefit from a zero trust architecture.

In the use cases below, ZTA is not explicitly indicated since the enterprise likely has both perimeter-based and possibly ZTA infrastructures. As discussed in Section 7.2, there will likely be a period when ZTA components and perimeter-based network infrastructure are concurrently in operation in an enterprise.

4.1 Enterprise with Satellite Facilities

The most common scenario involves an enterprise with a single headquarters and one or more geographically dispersed locations that are not joined by an enterprise-owned physical network connection (see Figure 8). Employees at the remote location may not have a full enterprise-owned local network but still need to access enterprise resources to perform their tasks. The enterprise may have a Multiprotocol Label Switch (MPLS) link to the enterprise HQ network but may not have adequate bandwidth for all traffic or may not wish for traffic destined for cloud-based applications/services to traverse through the enterprise HQ network. Likewise, employees may be teleworking or in a remote location and using enterprise-owned or personally-owned devices. In such cases, an enterprise may wish to grant access to some resources (e.g., employee calendar, email) but deny access or restrict actions to more sensitive resources (e.g., HR database).

In this use case, the PE/PA(s) is often hosted as a cloud service (which usually provides superior availability and would not require remote workers to rely on enterprise infrastructure to access cloud resources) with end assets having an installed agent (see Section 3.2.1) or accessing a resource portal (see Section 3.2.3). It may not be most responsive to have the PE/PA(s) hosted on the enterprise local network as remote offices and workers must send all traffic back to the enterprise network to reach applications/services hosted by cloud services.

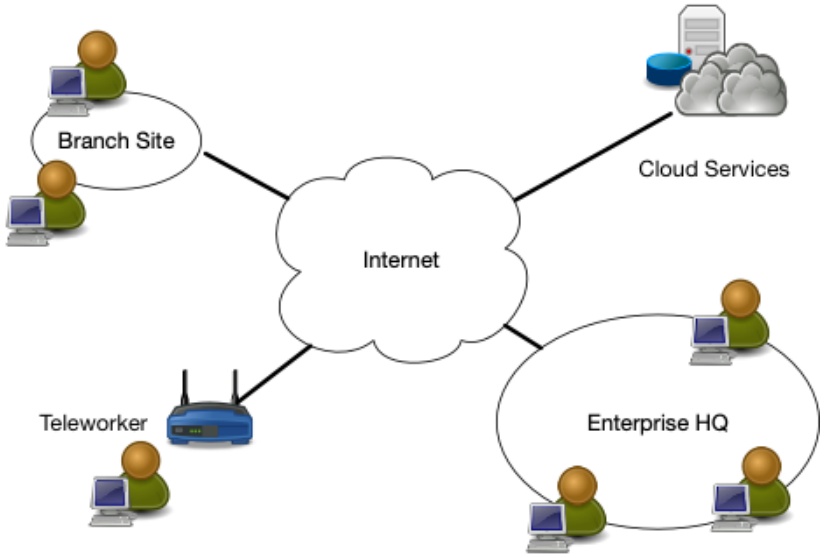

Figure 8: Enterprise with Remote Employees

4.2 Multi-cloud/Cloud-to-Cloud Enterprise

One increasingly common use case for deploying a ZTA is an enterprise utilizing multiple cloud providers (see Figure 9). In this use case, the enterprise has a local network but uses two or more cloud service providers to host applications/services and data. Sometimes, the application/service is hosted on a cloud service that is separate from the data source. For performance and ease of management, the application hosted in Cloud Provider A should be able to connect directly to the data source hosted in Cloud Provider B rather than force the application to tunnel back through the enterprise network.

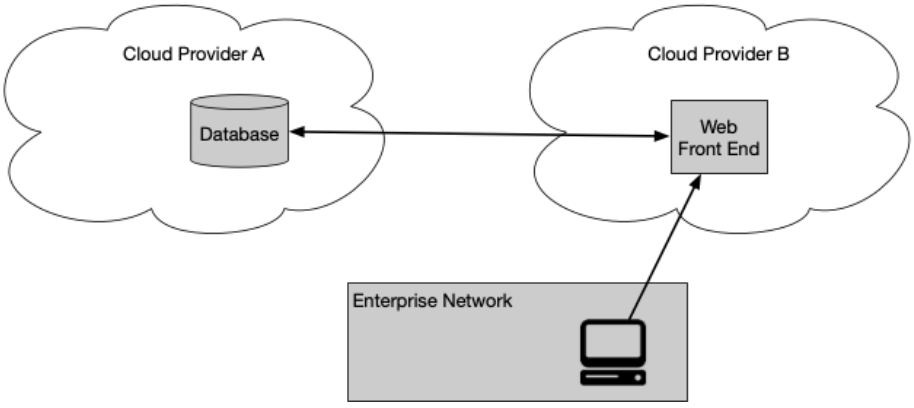

Figure 9: Multi-cloud Use Case

This use case is the server-server implementation of the CSA's software defined perimeter (SDP) specification [CSA-SDP]. As enterprises move to more cloud-hosted applications and services, it becomes apparent that relying on the enterprise perimeter for security becomes a liability. As discussed in Section 2.2, ZT principles take the view that there should be no difference between enterprise-owned and -operated network infrastructure and infrastructure owned and operated by any other service provider. The zero trust approach to multi-cloud use is to place PEPs at the

access points of each application/service and data source. The PE and PA may be services located in either cloud or even on a third cloud provider. The client (via a portal or local installed agent) then accesses the PEPs directly. That way, the enterprise can still manage access to resources even when hosted outside the enterprise. One challenge is that different cloud providers have unique ways of implementing similar functionality. Enterprise architects will need to be aware of the how to implement their enterprise ZTA with each cloud provider they utilize.

4.3 Enterprise with Contracted Services and/or Nonemployee Access

Another common scenario is an enterprise that includes on-site visitors and/or contracted service providers that require limited access to enterprise resources to do their work (see Figure 10). For example, an enterprise has its own internal applications/services, databases, and assets. These include services contracted out to providers who may occasionally be on-site to provide maintenance (e.g., smart heating and lighting systems that are owned and managed by external providers). These visitors and service providers will need network connectivity to perform their tasks. A zero trust enterprise could facilitate this by allowing these devices and any visiting service technician access to the internet while obscuring enterprise resources.

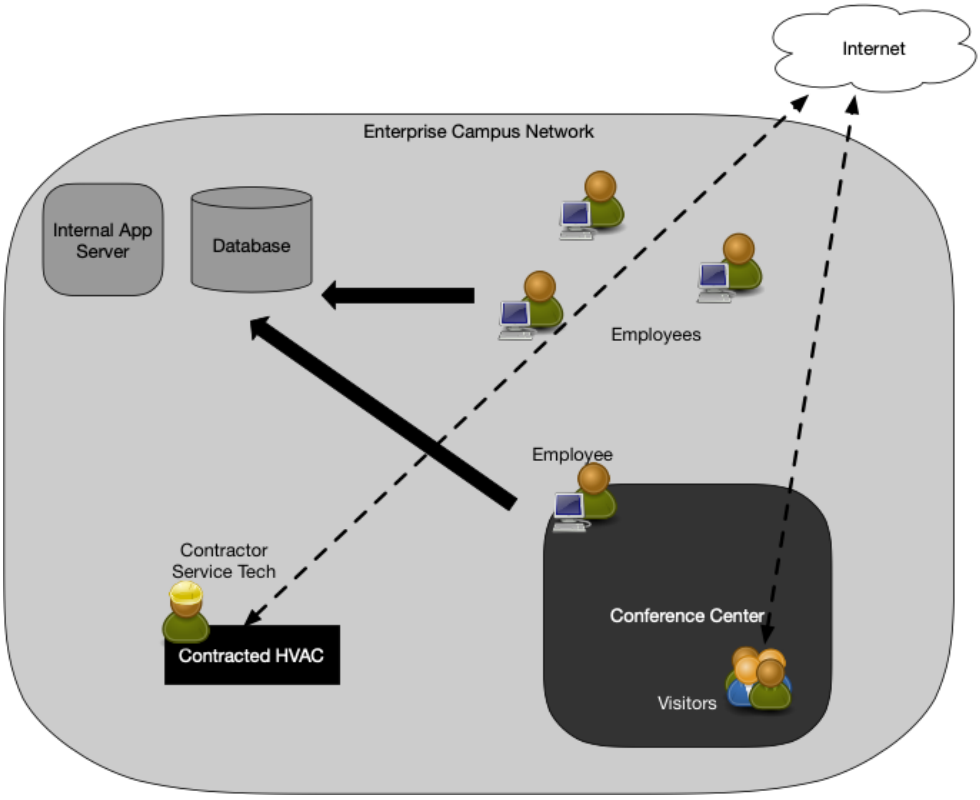

Figure 10: Enterprise with Nonemployee Access

In this example, the organization also has a conference center where visitors interact with employees. Again, with a ZTA approach of SDPs, employee devices and subjects are differentiated and may be able to access appropriate enterprise resources. Visitors to the campus can have internet access but cannot access enterprise resources. They may not even be able to

discover enterprise services via network scans (i.e., prevent active network reconnaissance/east-west movement).

In this use case, the PE(s) and PA(s) could be hosted as a cloud service or on the LAN (assuming little or no use of cloud-hosted services). The enterprise assets could have an installed agent (see Section 3.2.1) or access resources via a portal (see Section 3.2.3). The PA(s) ensures that all nonenterprise assets (those that do not have installed agents or cannot connect to a portal) cannot access local resources but may access the internet.

4.4 Collaboration Across Enterprise Boundaries

A fourth use case is cross-enterprise collaboration. For example, there is a project involving employees from Enterprise A and Enterprise B (see Figure 11). The two enterprises may be separate federal agencies (G2G) or even a federal agency and a private enterprise (G2B). Enterprise A operates the database used for the project but must allow access to the data for certain members of Enterprise B. Enterprise A can set up specialized accounts for the employees of Enterprise B to access the required data and deny access to all other resources, but this can quickly become difficult to manage. Having both organizations enrolled in a federated ID management system would allow quicker establishment of these relationships provided that both organizations' PEPs can authenticate subjects in a federated ID community.

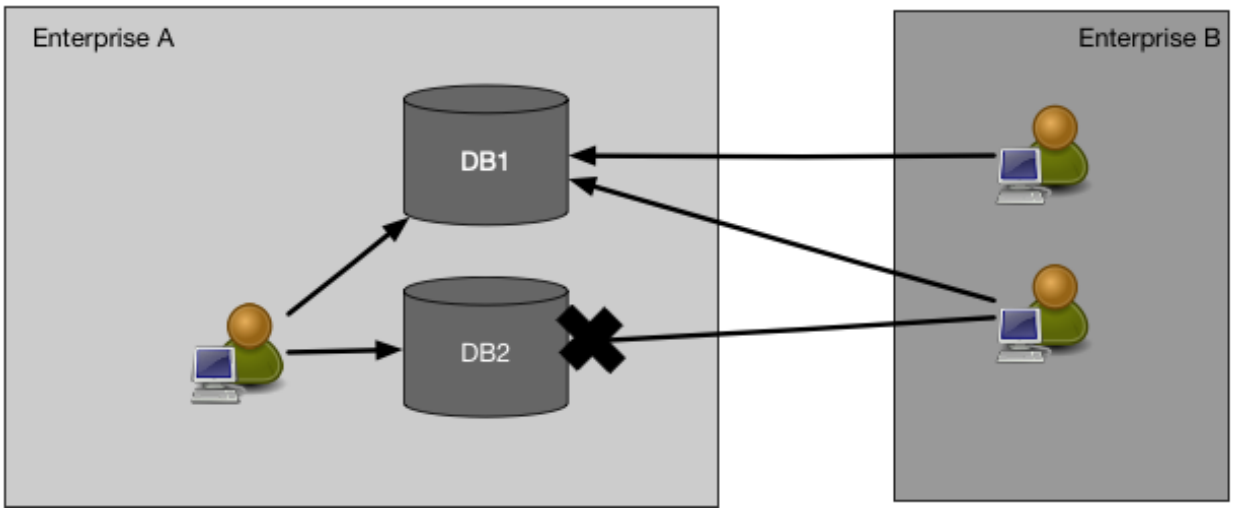

Figure 11: Cross-Enterprise Collaboration

This scenario can be similar to Use Case 1 (Section 4.1) as employees of both enterprises may not be located on their organizations' network infrastructures, and the resource they need to access may be within one enterprise environment or hosted in the cloud. This means that there do not need to be complex firewall rules or enterprise-wide access control lists (ACLs) allowing certain IP addresses belonging to Enterprise B to access resources in Enterprise A based on Enterprise A's access policies. How this access is accomplished depends on the technology in use. Similar to Use Case 1, a PE and PA hosted as a cloud service may provide availability to all parties without having to establish a VPN or similar. The employees of Enterprise B may be asked to install a software agent on their asset or access the necessary data resources through a web gateway (see Section 3.2.3).

4.5 Enterprise with Public- or Customer-Facing Services

A common feature in many enterprises is a public-facing service that may or may not include user registration (i.e., users must create or have been issued a set of login credentials). Such services could be for the general public, a set of customers with an existing business relationship, or a special set of nonenterprise users such as employee dependents. In all cases, it is likely that requesting assets are not enterprise-owned, and the enterprise is constrained as to what internal cybersecurity polices can be enforced.

For a general, public-facing resource that does not require login credentials to access (e.g., public web page), the tenets of ZTA do not directly apply. The enterprise cannot strictly control the state of requesting assets, and anonymous public resources (e.g., a public web page) do not require credentials in order to be accessed.

Enterprises may establish policies for registered public users such as customers (i.e., those with a business relationship) and special users (e.g., employee dependents). If the users are required to produce or are issued credentials, the enterprise may institute policies regarding password length, life cycle, and other details and may provide MFA as an option or requirement. However, enterprises are limited in the policies they can implement for this class of user. Information about incoming requests may be useful in determining the state of the public service and detecting possible attacks masquerading as legitimate users. For example, a registered user portal is known to be accessed by registered customers using one of a set of common web browsers. A sudden increase in access requests from unknown browser types or known outdated versions could indicate an automated attack of some kind, and the enterprise could take steps to limit requests from these identified clients. The enterprise should also be aware of any statutes or regulations regarding what information can be collected and recorded about the requesting users and assets.

5 Threats Associated with Zero Trust Architecture

No enterprise can eliminate cybersecurity risk. When complemented with existing cybersecurity policies and guidance, identity and access management, continuous monitoring, and general cyber hygiene, a properly implemented and maintained ZTA can reduce overall risk and protect against common threats. However, some threats have unique features when implementing a ZTA.

5.1 Subversion of ZTA Decision Process

In ZTA, the policy engine and policy administrator are the key components of the entire enterprise. No communication between enterprise resources occurs unless it is approved and possibly configured by the PE and PA. This means that these components must be properly configured and maintained. Any enterprise administrator with configuration access to the PE's rules may be able to perform unapproved changes or make mistakes that can disrupt enterprise operations. Likewise, a compromised PA could allow access to resources that would otherwise not be approved (e.g., to a subverted, personally-owned device). Mitigating associated risks means the PE and PA components must be properly configured and monitored, and any configuration changes must be logged and subject to audit.

5.2 Denial-of-Service or Network Disruption

In ZTA, the PA is the key component for resource access. Enterprise resources cannot connect to each other without the PA's permission and, possibly, configuration action. If an attacker disrupts or denies access to the PEP(s) or PE/PA (i.e., DoS attack or route hijack), it can adversely impact enterprise operations. Enterprises can mitigate this threat by having the policy enforcement reside in a properly secured cloud environment or be replicated in several locations following guidance on cyber resiliency [SP 800-160v2].

This mitigates the risk but does not eliminate it. Botnets such as Mirai produce massive DoS attacks against key internet service providers and disrupt service to millions of internet users.[5] It is also possible that an attacker could intercept and block traffic to a PEP or PA from a portion or all of the user accounts within an enterprise (e.g., a branch office or even a single remote employee). In such cases, only a portion of enterprise subjects is affected. This is also possible in legacy remote-access VPNs and is not unique to ZTA.

A hosting provider may also accidentally take a cloud-based PE or PA offline. Cloud services have experienced disruptions in the past, both infrastructure as a service (IaaS)[6] and SaaS.[7] An operational error could prevent an entire enterprise from functioning if the policy engine or policy administrator component becomes inaccessible from the network.

[5] https://blog.cloudflare.com/inside-mirai-the-infamous-iot-botnet-a-retrospective-analysis/
[6] https://aws.amazon.com/message/41926/
[7] https://www.nzherald.co.nz/business/news/article.cfm?c_id=3&objectid=12286870

There is also the risk that enterprise resources may not be reachable from the PA, so even if access is granted to a subject, the PA cannot configure the communication path from the network. This could happen due to a DDoS attack or simply due to unexpected heavy usage. This is similar to any other network disruption in that some or all enterprise subjects cannot access a particular resource due to that resource not being available for some reason.

5.3 Stolen Credentials/Insider Threat

Properly implemented ZT, information security and resiliency policies, and best practices reduce the risk of an attacker gaining broad access via stolen credentials or insider attack. The ZT principle of no implicit trust based on network location means attackers need to compromise an existing account or device to gain a foothold in an enterprise. A properly developed and implemented ZTA should prevent a compromised account or asset from accessing resources outside its normal purview or access patterns. This means that accounts with access policies around resources that an attacker is interested in would be the primary targets for attackers.

Attackers may use phishing, social engineering, or a combination of attacks to obtain credentials of valuable accounts. “Valuable” may mean different things based on the attacker’s motivation. For instance, enterprise administrator accounts may be valuable, but attackers interested in financial gain may consider accounts that have access to financial or payment resources of equal value. Implementation of MFA for access requests may reduce the risk of information loss from a compromised account. However, an attacker with valid credentials (or a malicious insider) may still be able to access resources for which the account has been granted access. For example, an attacker or compromised employee who has the credentials and enterprise-owned asset of a valid human resources employee may still be able to access an employee database.

ZTA reduces risk and prevents any compromised accounts or assets from moving laterally throughout the network. If the compromised credentials are not authorized to access a particular resource, they will continue to be denied access to that resource. In addition, a contextual trust algorithm (see Section 3.3.1) is more likely to detect and respond quickly to this attack than when occurring in a legacy, perimeter-based network. The contextual TA can detect access patterns that are out of normal behavior and deny the compromised account or insider threat access to sensitive resources.

5.4 Visibility on the Network

As mentioned in Section 3.4.1, all traffic is inspected and logged on the network and analyzed to identify and react to potential attacks against the enterprise. However, as also mentioned, some (possibly the majority) of the traffic on the enterprise network may be opaque to layer 3 network analysis tools. This traffic may originate from nonenterprise-owned assets (e.g., contracted services that use the enterprise infrastructure to access the internet) or applications/services that are resistant to passive monitoring. The enterprise that cannot perform deep packet inspection or examine the encrypted traffic and must use other methods to assess a possible attacker on the network.

That does not mean that the enterprise is unable to analyze encrypted traffic that it sees on the network. The enterprise can collect metadata (e.g., source and destination addresses, etc.) about

the encrypted traffic and use that to detect an active attacker or possible malware communicating on the network. Machine learning techniques [Anderson] can be used to analyze traffic that cannot be decrypted and examined. Employing this type of machine learning would allow the enterprise to categorize traffic as valid or possibly malicious and subject to remediation.

5.5 Storage of System and Network Information

A related threat to enterprise monitoring and analysis of network traffic is the analysis component itself. If monitor scans, network traffic, and metadata are being stored for building contextual policies, forensics, or later analysis, that data becomes a target for attackers. Just like network diagrams, configuration files, and other assorted network architecture documents, these resources should be protected. If an attacker can successfully gain access to this information, they may be able to gain insight into the enterprise architecture and identify assets for further reconnaissance and attack.

Another source of reconnaissance information for an attacker in a ZT enterprise is the management tool used to encode access policies. Like stored traffic, this component contains access policies to resources and can give an attacker information on which accounts are most valuable to compromise (e.g., the ones that have access to the desired data resources).

As for all valuable enterprise data, adequate protections should be in place to prevent unauthorized access and access attempts. As these resources are vital to security, they should have the most restrictive access policies and be accessible only from designated or dedicated administrator accounts.

5.6 Reliance on Proprietary Data Formats or Solutions

ZTA relies on several different data sources to make access decisions, including information about the requesting subject, asset used, enterprise and external intelligence, and threat analysis. Often, the assets used to store and process this information do not have a common, open standard on how to interact and exchange information. This can lead to instances where an enterprise is locked into a subset of providers due to interoperability issues. If one provider has a security issue or disruption, an enterprise may not be able to migrate to a new provider without extreme cost (e.g., replacing several assets) or going through a long transition program (e.g., translating policy rules from one proprietary format to another). Like DoS attacks, this risk is not unique to ZTA, but because ZTA is heavily dependent on the dynamic access of information (both enterprise and service providers), disruption can affect the core business functions of an enterprise. To mitigate associated risks, enterprises should evaluate service providers on a holistic basis by considering factors such as vendor security controls, enterprise switching costs, and supply chain risk management in addition to more typical factors such as performance, stability, etc.

5.7 Use of Non-person Entities (NPE) in ZTA Administration

Artificial intelligence and other software-based agents are being deployed to manage security issues on enterprise networks. These components need to interact with the management components of ZTA (e.g., policy engine, policy administrator), sometimes in lieu of a human administrator. How these components authenticate themselves in an enterprise implementing a

ZTA is an open issue. It is assumed that most automated technology systems will use some means to authenticate when using an API to resource components.

The biggest risk when using automated technology for configuration and policy enforcement is the possibility of false positives (innocuous actions mistaken for attacks) and false negatives (attacks mistaken for normal activity) impacting the security posture of the enterprise. This can be reduced with regular retuning analysis to correct mistaken decisions and improve the decision process.

The associated risk is that an attacker will be able to induce or coerce an NPE to perform some task that the attacker is not privileged to perform. The software agent may have a lower bar for authentication (e.g., API key versus MFA) to perform administrative or security-related tasks compared with a human user. If an attacker can interact with the agent, they could theoretically trick the agent into allowing the attacker greater access or into performing some task on behalf of the attacker. There is also a risk that an attacker could gain access to a software agent's credentials and impersonate the agent when performing tasks.

6 Zero Trust Architecture and Possible Interactions with Existing Federal Guidance

Several existing federal policies and guidance intersect with the planning, deployment, and operation of a ZTA. These policies do not prohibit an enterprise from moving to a more zero trust-oriented architecture but can influence development of a zero trust strategy for an agency. When complemented with existing cybersecurity policies and guidance, ICAM, continuous monitoring, and general cyber hygiene, ZTA may reinforce an organization's security posture and protect against common threats.

6.1 ZTA and NIST Risk Management Framework

A ZTA deployment involves developing access polices around acceptable risk to the designated mission or business process (see Section 7.3.3). It is possible to deny all network access to a resource and allow access only via a connected terminal, but this is disproportionately restrictive in the majority of cases and could inhibit work from being accomplished. For a federal agency to perform its mission, there is an acceptable level of risk. The risks associated with performing the given mission must be identified and evaluated, and either accepted or mitigated. To assist in this, the NIST Risk Management Framework (RMF) was developed [SP800-37].

ZTA planning and implementation may change the authorization boundaries defined by the enterprise. This is due to the addition of new components (e.g., policy engine, policy administrator, and PEPs) and a reduction of reliance on network perimeter defenses. The overall process described in the RMF will not change in a ZTA.

6.2 Zero Trust and NIST Privacy Framework

Protecting the privacy of users and private information (e.g., personally identifiable information) is a prime concern for organizations. Privacy and data protections are included in compliance programs such as FISMA and the Heath Insurance Portability and Accountability Act (HIPAA). In response, NIST produced a Privacy Framework for use by organizations [NISTPRIV]. This document provides a framework to describe privacy risks and mitigation strategies, as well as a process for an enterprise to identify, measure, and mitigate risks to user privacy and private information stored and processed by an organization. This includes personal information used by the enterprise to support ZTA operations and any biometric attributes used in access request evaluations.

Part of the core requirements for ZT is that an enterprise should inspect and log traffic (or at least log and inspect metadata when dealing with traffic that cannot be decrypted by monitoring systems) in its environment. Some of this traffic may contain private information or have associated privacy risks. Organizations will need to identify any possible risks associated with intercepting, scanning, and logging network traffic [NISTIR 8062]. This may include actions such as informing users, obtaining consent (via a login page, banner, or similar), and educating enterprise users. The NIST Privacy Framework [NISTPRIV] could help in developing a formal process to identify and mitigate any privacy-related risks to an enterprise developing a zero trust architecture.

6.3 ZTA and Federal Identity, Credential, and Access Management Architecture

Subject provisioning is a key component of ZTA. The policy engine cannot determine if attempted connections are authorized to connect to a resource if the PE has insufficient information to identify associated subjects and resources. Strong subject provision and authentication policies need to be in place before moving to a more zero trust–aligned deployment. Enterprises need a clear set of subject attributes and policies that can be used by a PE to evaluate access requests.

The Office of Management and Budget (OMB) issued M-19-17 on improving identity management for the Federal Government. The goal of the policy is to develop "…a common vision for identity as an enabler of mission delivery, trust, and safety of the Nation" [M-19-17]. The memo calls on all federal agencies to form an ICAM office to govern efforts related to identity issuance and management. Many of these management policies should use the recommendations in NIST SP 800-63-3, *Digital Identity Guidelines* [SP800-63]. As ZTA is heavily dependent on precise identity management, any ZTA effort will need to integrate the agency's ICAM policy.

6.4 ZTA and Trusted Internet Connections 3.0

TIC is a federal cybersecurity initiative jointly managed by OMB, DHS, and the General Services Administration (GSA), and is intended to establish a network security baseline across the Federal Government. Historically, TIC was a perimeter-based cybersecurity strategy which required agencies to consolidate and monitor their external network connections. Inherent in TIC 1.0 and TIC 2.0 is the assumption that the inside of the perimeter is "trusted," whereas ZTA assumes that network location does not infer "trust" (i.e., there is no "trust" on an agency's internal network). TIC 2.0 provides a list of network-based security capabilities (e.g. content filtering, monitoring, authentication, and others) to be deployed at the TIC Access Point at the agency's perimeter; many of these capabilities are aligned with ZT principles.

TIC 3.0 has been updated to accommodate cloud services and mobile devices [M-19-26]. In TIC 3.0, it is recognized that the definition of "trust" may vary across specific computing contexts and that agencies have different risk tolerances for defining trust zones. In addition, TIC 3.0 has an updated TIC Security Capability Handbook, which defines two types of security capabilities: (1) Universal Security Capabilities that apply at the enterprise level, and (2) PEP Security Capabilities that are network-level capabilities to be applied to multiple policy enforcement points (PEPs), as defined in TIC use cases. The PEP Security Capabilities may be applied at any appropriate PEP located along a given data flow instead of at a single PEP at the agency perimeter. Many of these TIC 3.0 security capabilities directly support ZTA (e.g., encrypted traffic, strong authentication, microsegmentation, network and system inventory, and others). TIC 3.0 defines specific use cases that describe the implementation of trust zones and security capabilities across specific applications, services, and environments.

TIC 3.0 is focused on network-based security protections, whereas ZTA is a more inclusive architecture addressing application, user, and data protections. As TIC 3.0 evolves its use cases, it is likely that a ZTA TIC use case will be developed to define the network protections to be deployed at ZTA enforcement points.

6.5 ZTA and EINSTEIN (NCPS – National Cybersecurity Protection System)

NCPS (operationally known as EINSTEIN) is an integrated system-of-systems that delivers intrusion detection, advanced analytics, information sharing, and intrusion prevention capabilities to defend the Federal Government from cyber threats. The goals of NCPS, which align with the overarching goals of zero trust, are to manage cyber risk, improve cyber protection, and empower partners to secure cyber space. EINSTEIN sensors enable CISA's National Cybersecurity and Communications Integration Center (NCCIC) to defend federal networks and respond to significant incidents at federal agencies.

The placement of NCPS sensors for DHS situational awareness is based on a perimeter network defense in the Federal Government, while ZTA moves protections closer to the assets, data and all other resources. The NCPS program is evolving to ensure that situational awareness is preserved through utilization of security information about cloud-based traffic, helping to set the foundation for expanded situational awareness telemetry from ZTA systems. NCPS intrusion prevention functions would also require evolution to be able to inform policy enforcement at both the current NCPS locations as well as ZTA systems. As ZTA is adopted across the Federal Government, the NCPS implementation would need to continually evolve, or new capabilities would need to be deployed to fulfill NCPS objectives. Incident responders could potentially leverage information from the authentication, traffic inspection, and logging of agency traffic available to federal agencies that have implemented a zero trust architecture. Information generated in a ZTA may better inform event impact quantification; machine learning tools could use ZTA data to improve detection; and additional logs from ZTA may be saved for after-the-fact analyses by incident responders.

6.6 ZTA and DHS Continuous Diagnostics and Mitigations (CDM) Program

The DHS CDM program is an effort to improve federal agency information technology (IT). Vital to that posture is an agency's insight into the assets, configuration, and subjects within itself. To protect a system, agencies need to set up processes to discover and understand the basic components and actors in their infrastructure:

- **What is connected?** What devices, applications, and services are used by the organization? This includes observing and improving the security posture of these artifacts as vulnerabilities and threats are discovered.
- **Who is using the network?** What users are part of the organization or are external and allowed to access enterprise resources? These include NPEs that may be performing autonomous actions.
- **What is happening on the network?** An enterprise needs insight into traffic patterns and messages between systems.
- **How is data protected?** The enterprise needs a set policy on how information is protected at rest, in transit, and in use.

Having a strong CDM program implementation is key to the success of ZTA. For example, to move to ZTA, an enterprise must have a system to discover and record physical and virtual assets to create a usable inventory. The DHS CDM program has initiated several efforts to build the capabilities needed within federal agencies to move to a ZTA. For example, the DHS

Hardware Asset Management (HWAM) [HWAM] program is an effort to help agencies identify devices on their network infrastructure to deploy a secure configuration. This is similar to the first steps in developing a road map to ZTA. Agencies must have visibility into the assets active on the network (or those accessing resources remotely) to categorize, configure, and monitor the network's activity.

6.7 ZTA, Cloud Smart, and the Federal Data Strategy

The Cloud Smart[8] strategy, updated Data Center Optimization Initiative [M-19-19] policy, and Federal Data Strategy[9] all influence some requirements for agencies planning a ZTA. These policies require agencies to inventory and assess how they collect, store, and access data, both on premises and in the cloud.

This inventory is critical to determining what business processes and resources would benefit from implementing a ZTA. Data resources and applications and services that are primarily cloud-based or primarily used by remote workers are good candidates for a ZTA approach (see Section 7.3.3) because the subjects and resources are located outside of the enterprise network perimeter and are likely to see the most benefit in use, scalability, and security.

One additional consideration with the Federal Data Strategy is how to make agency data assets accessible to other agencies or the public. This corresponds with the cross-enterprise collaboration ZTA use case (see Section 4.4). Agencies using a ZTA for these assets may need to take collaboration or publication requirements into account when developing the strategy.

[8] Federal Cloud Computing Strategy: https://cloud.cio.gov/strategy/
[9] Federal Data Strategy: https://strategy.data.gov/

7 Migrating to a Zero Trust Architecture

Implementing a ZTA is a journey rather than a wholesale replacement of infrastructure or processes. An organization should seek to incrementally implement zero trust principles, process changes, and technology solutions that protect its highest value data assets. Most enterprises will continue to operate in a hybrid zero-trust/perimeter-based mode for an indefinite period while continuing to invest in ongoing IT modernization initiatives. Having an IT modernization plan that includes moving to an architecture based on ZT principles may help an enterprise form roadmaps for small scale workflow migrations.

How an enterprise migrates to a strategy depends on its current cybersecurity posture and operations. An enterprise should reach a baseline of competence before it becomes possible to deploy a significant ZT-focused environment [ACT-IAC]. This baseline includes having assets, subjects, business processes, traffic flows and dependency mappings identified and cataloged for the enterprise. The enterprise needs this information before it can develop a list of candidate business processes and the subjects/assets involved in this process.

7.1 Pure Zero Trust Architecture

In a greenfield approach, it would be possible to build a zero trust architecture from the ground up. Assuming the enterprise knows the applications/services and workflows that it wants to use for its operations, it can produce an architecture based on zero trust tenets for those workflows. Once the workflows are identified, the enterprise can narrow down the components needed and begin to map how the individual components interact. From that point, it is an engineering and organizational exercise in building the infrastructure and configuring the components. This may include additional organizational changes depending on how the enterprise is currently set up and operating.

In practice, this is rarely a viable option for federal agencies or any organization with an existing network. However, there may be times when an organization is asked to fulfill a new responsibility that would require building its own infrastructure. In these cases, it might be possible to introduce ZT concepts to some degree. For example, an agency may be given a new responsibility that entails building a new application, service, or database. The agency could design the newly needed infrastructure around ZT principles and secure system engineering [SP8900-160v1], such as evaluating subjects' trust before granting access and establishing micro-perimeters around new resources. The degree of success depends on how dependent this new infrastructure is on existing resources (e.g., ID management systems).

7.2 Hybrid ZTA and Perimeter-Based Architecture

It is unlikely that any significant enterprise can migrate to zero trust in a single technology refresh cycle. There may be an indefinite period when ZTA workflows coexist with non-ZTA workflows in an enterprise. Migration to a ZTA approach to the enterprise may take place one business process at a time. The enterprise needs to make sure that the common elements (e.g., ID management, device management, event logging) are flexible enough to operate in a ZTA and perimeter-based hybrid security architecture. Enterprise architects may also want to restrict ZTA candidate solutions to those that can interface with existing components.

Migrating an existing workflow to a ZTA will likely require (at least) a partial redesign. Enterprises may take this opportunity to adopt secure system engineering [SP800-160v1] practices if they have not already done so for workflows.

7.3 Steps to Introducing ZTA to a Perimeter-Based Architected Network

Migrating to ZTA requires an organization to have detailed knowledge of its assets (physical and virtual), subjects (including user privileges), and business processes. This knowledge is accessed by the PE when evaluating resource requests. Incomplete knowledge will most often lead to a business process failure where the PE denies requests due to insufficient information. This is especially an issue if there are unknown "shadow IT" deployments operating within an organization.

Before undertaking an effort to bring ZTA to an enterprise, there should be a survey of assets, subjects, data flows, and workflows. This awareness forms the foundational state that must be reached before a ZTA deployment is possible. An enterprise cannot determine what new processes or systems need to be in place if there is no knowledge of the current state of operations. These surveys can be conducted in parallel, but both are tied to examination of the business processes of the organization. These steps can be mapped to the steps in the RMF [SP800-37] as any adoption of a ZTA is a process to reduce risk to an agency's business functions. The pathway to implementing a ZTA can be visualized in Figure 12.

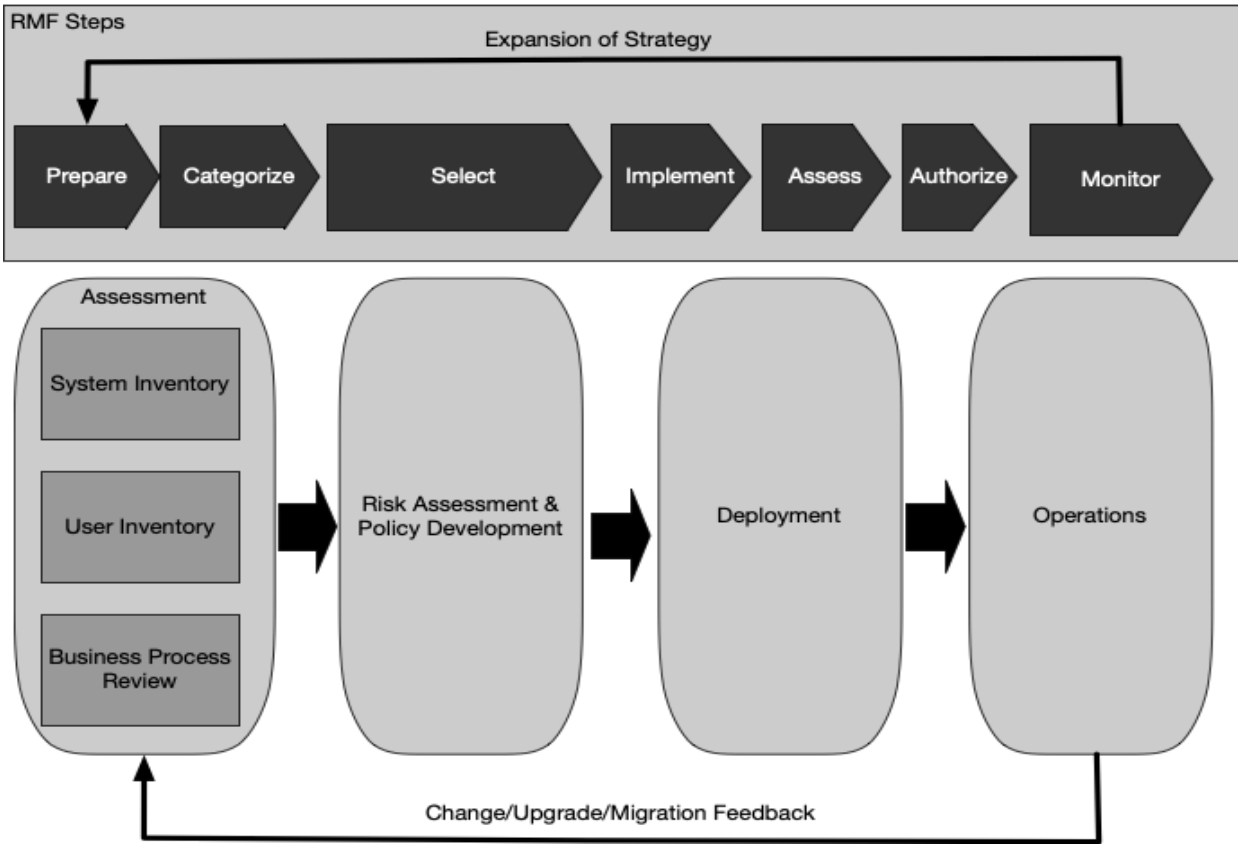

Figure 12: ZTA Deployment Cycle

After the initial inventory is created, there is a regular cycle of maintenance and updating. This updating may either change business processes or not have any impact, but an evaluation of business processes should be conducted. For example, a change in digital certificate providers may not appear to have a significant impact but may involve certificate root store management, Certificate Transparency log monitoring, and other factors that are not apparent at first.

7.3.1 Identify Actors on the Enterprise

For a zero trust enterprise to operate, the PE must have knowledge of enterprise subjects. Subjects could encompass both human and possible NPEs, such as service accounts that interact with resources.

Users with special privileges, such as developers or system administrators, require additional scrutiny when being assigned attributes or roles. In many legacy security architectures, these accounts may have blanket permission to access all enterprise resources. ZTA should allow developers and administrators to have sufficient flexibility to satisfy their business requirements while using logs and audit actions to identify access behavior patterns. ZTA deployments may require administrators to satisfy a more stringent confidence level or criteria as outlined in NIST SP 800-63A, Section 5 [SP800-63A].

7.3.2 Identify Assets Owned by the Enterprise

As mentioned in Section 2.1, one of the key requirements of ZTA is the ability to identify and manage devices. ZTA also requires the ability to identify and monitor nonenterprise-owned devices that may be on enterprise-owned network infrastructure or that access enterprise resources. The ability to manage enterprise assets is key to the successful deployment of ZTA. This includes hardware components (e.g., laptops, phones, IoT devices) and digital artifacts (e.g., user accounts, applications, digital certificates). It may not be possible to conduct a complete census on all enterprise-owned assets, so an enterprise should consider building the capability to quickly identify, categorize, and assess newly discovered assets that are on enterprise-owned infrastructure.

This goes beyond simply cataloging and maintaining a database of enterprise assets. This also includes configuration management and monitoring. The ability to observe the current state of an asset is part of the process of evaluating access requests (see Section 2.1). This means that the enterprise must be able to configure, survey, and update enterprise assets, such as virtual assets and containers. This also includes both its physical (as best estimated) and network location. This information should inform the PE when making resource access decisions.

Nonenterprise-owned assets and enterprise-owned “shadow IT” should also be cataloged as well as possible. This may include whatever is visible by the enterprise (e.g., MAC address, network location) and augmented by administrator data entry. This information is not only used for access decisions (as collaborator and BYOD assets may need to contact PEPs) but also for monitoring and forensics logging by the enterprise. Shadow IT presents a special problem in that these resources are enterprise-owned but not managed like other resources. Certain ZTA approaches (mainly network-based) may even cause shadow IT components to become unusable as they may not be known and included in network access policies.

Many federal agencies have already begun identifying enterprise assets. Agencies that have established CDM program capabilities, such as HWAM [HWAM] and Software Asset Management (SWAM) [SWAM], have a rich set of data to draw from when enacting a ZTA. Agencies may also have a list of ZTA candidate processes that involve High Value Assets (HVA) [M-19-03] that have been identified as key to the agency mission. This work would need to exist enterprise- or agency-wide before any business process could be (re)designed with a ZTA. These programs must be designed to be expandable and adaptable to changes in the enterprise, not only when migrating to ZTA but also when accounting for new assets, services, and business processes that become part of the enterprise.

7.3.3 Identify Key Processes and Evaluate Risks Associated with Executing Process

The third inventory that an agency should undertake is to identify and rank the business processes, data flows, and their relation in the missions of the agency. Business processes should inform the circumstances under which resource access requests are granted and denied. An enterprise may wish to start with a low-risk business process for the first transition to ZTA as disruptions will likely not negatively impact the entire organization. Once enough experience is gained, more critical business processes can become candidates.

Business processes that utilize cloud-based resources or are used by remote workers are often good candidates for ZTA and would likely see improvements to availability and security. Rather than project the enterprise perimeter into the cloud or bring clients into the enterprise network via a VPN, enterprise clients can request cloud services directly. The enterprise's PEPs ensure that enterprise policies are followed before resource access is granted to a client. Planners should also consider potential tradeoffs in performance, user experience, and possible increased workflow fragility that may occur when implementing ZTA for a given business process.

7.3.4 Formulating Policies for the ZTA Candidate

The process of identifying a candidate service or business workflow depends on several factors: the importance of the process to the organization, the group of subjects affected, and the current state of resources used for the workflow. The value of the asset or workflow based on risk to the asset or workflow can be evaluated using the NIST Risk Management Framework [SP800-37].

After the asset or workflow is identified, identify all upstream resources (e.g., ID management systems, databases, micro-services), downstream resources (e.g., logging, security monitoring), and entities (e.g., subjects, service accounts) that are used or affected by the workflow. This may influence the candidate choice as a first migration to ZTA. An application/service used by an identified subset of enterprise subjects (e.g., a purchasing system) may be preferred over one that is vital to the entire subject base of the enterprise (e.g., email).

The enterprise administrators then need to determine the set of criteria (if using a criteria-based TA) or confidence level weights (if using a score-based TA) for the resources used in the candidate business process (see Section 3.3.1). Administrators may need to adjust these criteria or values during the tuning phase. These adjustments are necessary to ensure that policies are effective but do not hinder access to resources.

7.3.5 Identifying Candidate Solutions

Once a list of candidate business processes has been developed, enterprise architects can compose a list of candidate solutions. Some deployment models (see Section 3.1) are better suited to particular workflows and current enterprise ecosystems. Likewise, some vendor solutions are better suited to some use cases than others. These are some factors to consider:

- **Does the solution require that components be installed on the client asset?** This may limit business processes where nonenterprise-owned assets are used or desired, such as BYOD or cross-agency collaborations.
- **Does the solution work where the business process resources exist entirely on enterprise premises?** Some solutions assume that requested resources will reside in the cloud (so-called north-south traffic) and not within an enterprise perimeter (east-west traffic). The location of candidate business process resources will influence candidate solutions as well as the ZTA for the process.
- **Does the solution provide a means to log interactions for analysis?** A key component of ZT is the collection and use of data related to the process flow that feeds back into the PE when making access decisions.
- **Does the solution provide broad support for different applications, services, and protocols?** Some solutions may support a broad range of protocols (web, secure shell [SSH], etc.) and transports (IPv4 and IPv6), while others may only work with a narrow focus such as web or email.
- **Does the solution require changes to subject behavior?** Some solutions may require additional steps to perform a given workflow. This may change how enterprise subjects perform the workflow.

One solution is to model an existing business process as a pilot program rather than just a replacement. This pilot program could be made general to apply to several business processes or be made specific to one use case. The pilot can be used as a "proving ground" for ZTA before transitioning subjects to the ZTA deployment and away from the legacy process infrastructure.

7.3.6 Initial Deployment and Monitoring

Once the candidate workflow and ZTA components are chosen, the initial deployment can start. Enterprise administrators must implement the developed policies by using the selected components but may wish to operate in an observation and monitoring mode at first. Few enterprise policy sets are complete in their first iterations: important user accounts (e.g., administrator accounts) may be denied access to resources they need or may not need all the access privileges they have been assigned.

The new ZT business workflow could be operated in reporting-only mode for some time to make sure the policies are effective and workable. This also allows the enterprise to gain an understanding of baseline asset and resource access requests, behavior, and communication patterns. Reporting-only means that access should be granted for most requests, and logs and traces of connections should be compared with the initial developed policy. Basic policies such

as denying requests that fail MFA or appear from known, attacker controlled or subverted IP addresses should be enforced and logged, but after initial deployment, access polices should be more lenient to collect data from actual interactions of the ZT workflow. Once the baseline activity patterns for the workflow has been established, anomalous behavior can be more easily identify. If it is not possible to operate in a more lenient nature, enterprise network operators should monitor logs closely and be prepared to modify access policies based on operational experience.

7.3.7 Expanding the ZTA

When enough confidence is gained and the workflow policy set is refined, the enterprise enters the steady operational phase. The network and assets are still monitored, and traffic is logged (see Section 2.1), but responses and policy modifications are done at a lower tempo as they should not be severe. The subjects and stakeholders of the resources and processes involved should also provide feedback to improve operations. At this stage, the enterprise administrators can begin planning the next phase of ZT deployment. Like the previous rollout, a candidate workflow and solution set need to be identified and initial policies developed.

However, if a change occurs to the workflow, the operating ZT architecture needs to be reevaluated. Significate changes to the system—such as new devices, major updates to software (especially ZT logical components), and shifts in organizational structure—may result in changes to the workflow or policies. In effect, the entire process should be reconsidered with the assumption that some of the work has already been done. For example, new devices have been purchased, but no new user accounts have been created, so only the device inventory needs to be updated.

References

[ACT-IAC] American Council for Technology and Industry Advisory Council (2019) *Zero Trust Cybersecurity Current Trends*. Available at https://www.actiac.org/zero-trust-cybersecurity-current-trends

[Anderson] Anderson B, McGrew D (2017) Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Halifax, Nova Scotia, Canada), pp 1723-1732. https://doi.org/10.1145/3097983.3098163

[BCORE] Department of Defense CIO (2007). Department of Defense Global Information Grid Architecture Vision Version 1.0 June 2007. Available at http://www.acqnotes.com/Attachments/DoD%20GIG%20Architectural%20Vision,%20June%2007.pdf

[CSA-SDP] Cloud Security Alliance (2015) SDP Specification 1.0. Available at https://cloudsecurityalliance.org/artifacts/sdp-specification-v1-0/

[FIPS199] National Institute of Standards and Technology (2004) Standards for Security Categorization of Federal Information and Information Systems. (U.S. Department of Commerce, Washington, DC), Federal Information Processing Standards Publication (FIPS) 199. https://doi.org/10.6028/NIST.FIPS.199

[Gilman] Gilman E, Barth D (2017) *Zero Trust Networks: Building Secure Systems in Untrusted Networks* (O'Reilly Media, Inc., Sebastopol, CA), 1st Ed.

[HWAM] Department of Homeland Security (2015) *Hardware Asset Management (HWAM) Capability Description*. Available at https://www.us-cert.gov/sites/default/files/cdm_files/HWAM_CapabilityDescription.pdf

[IBNVN] Cohen R, Barabash K, Rochwerger B, Schour L, Crisan D, Birke R, Minkenberg C, Gusat M, Recio R, Jain V (2013) An Intent-based Approach for Network Virtualization. *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*. (IEEE, Ghent, Belgium), pp 42-50. Available at https://ieeexplore.ieee.org/document/6572968

[JERICHO] The Jericho Forum (2007) *Jericho Forum Commandments*, version 1.2. Available at https://collaboration.opengroup.org/jericho/commandments_v1.2.pdf

[M-19-03] Office of Management and Budget (2018) Strengthening the Cybersecurity of Federal Agencies by Enhancing the High Value Asset

Program. (The White House, Washington, DC), OMB Memorandum M-19-03, December 10, 2018. Available at https://www.whitehouse.gov/wp-content/uploads/2018/12/M-19-03.pdf

[M-19-17] Office of Management and Budget (2019) Enabling Mission Delivery through Improved Identity, Credential, and Access Management. (The White House, Washington, DC), OMB Memorandum M-19-17, May 21, 2019. Available at https://www.whitehouse.gov/wp-content/uploads/2019/05/M-19-17.pdf

[M-19-19] Office of Management and Budget (2019) Update on Data Center Optimization Initiative (DCOI). (The White House, Washington, DC), OMB Memorandum M-19-19, June 25, 2019. Available at https://datacenters.cio.gov/assets/files/m_19_19.pdf

[M-19-26] Office of Management and Budget (2019) Update to the Trusted Internet Connections (TIC) Initiative. (The White House, Washington, DC), OMB Memorandum M-19-26, September 12, 2019. Available at https://www.whitehouse.gov/wp-content/uploads/2019/09/M-19-26.pdf

[NISTIR 7987] Ferraiolo DF, Gavrila S, Jansen W (2015) Policy Machine: Features, Architecture, and Specification. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) 7987, Rev. 1. https://doi.org/10.6028/NIST.IR.7987r1

[NISTIR 8062] Brooks SW, Garcia ME, Lefkovitz NB, Lightman S, Nadeau EM (2017) An Introduction to Privacy Engineering and Risk Management in Federal Systems. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) 8062. https://doi.org/10.6028/NIST.IR.8062

[NISTPRIV] National Institute of Standards and Technology (2020) Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0. (National Institute of Standards and Technology, Gaithersburg, MD). https://doi.org/10.6028/NIST.CSWP.01162020

[SDNBOOK] Nadeau T, Gray K (2013) *SDN: Software Defined Networks: An Authoritative Review of Network Programmability Technologies*. (O'Reilly) 1st Ed.

[SP800-37] Joint Task Force (2018) 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), NIST Special Publication (SP) 800-37, Rev. 2. https://doi.org/10.6028/NIST.SP.800-37r2

[SP800-63] Grassi PA, Garcia ME, Fenton JL (2017) Digital Identity Guidelines. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-63-3, Includes updates as of March 2, 2020. https://doi.org/10.6028/NIST.SP.800-63-3

[SP800-63A] Grassi PA, Fenton JL, Lefkovitz NB, Danker JM, Choong Y-Y, Greene KK, Theofanos MF (2017) Digital Identity Guidelines: Enrollment and Identity Proofing. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-63A, Includes updates as of March 2, 2020. https://doi.org/10.6028/NIST.SP.800-63A

[SP800-160v1] Ross R, McEvilley M, Oren JC (2016) Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-160, Vol. 1, Includes updates as of March 21, 2018. https://doi.org/10.6028/NIST.SP.800-160v1

[SP800-160v2] Ross R, Pillitteri V, Graubart R, Bodeau D, McQuaid R (2019) Developing Cyber Resilient Systems: A Systems Security Engineering Approach. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-160, Vol. 2. https://doi.org/10.6028/NIST.SP.800-160v2

[SP800-162] Hu VC, Ferraiolo DF, Kuhn R, Schnitzer A, Sandlin K, Miller R, Scarfone KA (2014) Guide to Attribute Based Access Control (ABAC) Definition and Considerations. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-162, Includes updates as of August 2, 2019. https://doi.org/10.6028/NIST.SP.800-162

[SWAM] Department of Homeland Security (2015) *Software Asset Management (SWAM) Capability Description*. Available at https://www.us-cert.gov/sites/default/files/cdm_files/SWAM_CapabilityDescription.pdf

Appendix A—Acronyms

API	Application Programming Interface
BYOD	Bring Your Own Device
CDM	Continuous Diagnostics and Mitigation
DHS	Department of Homeland Security
DoS	Denial of Service
G2B	Government to Business (private industry)
G2G	Government to Government
NIST	National Institute of Standards and Technology
NPE	Non-Person Entity
PA	Policy Administrator
PDP	Policy Decision Point
PE	Policy Engine
PEP	Policy Enforcement Point
PKI	Public Key Infrastructure
RMF	NIST Risk Management Framework
SDN	Software Defined Network
SDP	Software Defined Perimeter
SIEM	Security Information and Event Monitoring
TIC	Trusted Internet Connections
VPN	Virtual Private Network
ZT	Zero Trust
ZTA	Zero Trust Architecture

Appendix B—Identified Gaps in the Current State-of-the-Art in ZTA

The current maturity of zero trust components and solutions was surveyed during the research conducted in the development of this document. This survey concluded that the current state of the ZTA ecosystem is not mature enough for widespread adoption. While it is possible to use ZTA strategies to plan and deploy an enterprise environment, there is no single solution that provides all the necessary components. Also, few ZTA components available today can be used for all of the various workflows present in an enterprise.

The following is a summary of identified gaps in the ZTA ecosystem and areas that need further investigation. Some of these areas have some foundation of work, but how ZTA tenets change these areas is not well-known as there is not enough experience with diverse ZTA-focused enterprise environments.

B.1 Technology Survey

Multiple vendors were invited to present their products and views on zero trust. The goal of this survey was to identify missing pieces that prevent agencies from moving to a zero trust based enterprise infrastructure now or maintaining an existing ZTA implementation. These gaps can be categorized into immediate deployment (immediate or short term), systemic gaps that affect maintenance or operations (short or midterm), and missing knowledge (areas for future research). They are summarized in Table B-1.

Table B-1: Summary of Identified Deployment Gaps

Category	Example Questions	Identified Gaps
Immediate deployment	• How should procurement requirements be written? • How does a ZTA plan work with TIC, FISMA, and other requirements?	• Lack of a common framework and vocabulary for ZTA • Perception that ZTA conflicts with existing policy
Systemic	• How can vendor lock-in be prevented? • How do different ZTA environments interact?	• Too much reliance on vendor APIs
Areas needing more research	• How will threats evolve in the face of ZTA? • How will business processes change in the face of ZTA?	• What a successful compromise looks like in an enterprise with a ZTA • Documented end user experience in an enterprise with a ZTA

B.2 Gaps that Prevent an Immediate Move to ZTA

These are the issues that are slowing adoption of a ZTA at present. These were classified as immediate issues, and no thought of future maintenance or migration was considered for this category. A forward-thinking enterprise may also consider the maintenance category to be of immediate concern in preventing the initial deployment of ZTA components, but these issues are considered a separate category for this analysis.

B.2.1 Lack of Common Terms for ZTA Design, Planning, and Procurement

Zero trust as a strategy for the design and deployment of enterprise infrastructure is still a forming concept. Industry has not yet coalesced around a single set of terms or concepts to describe ZTA components and operations. This makes it difficult for organizations (e.g., federal agencies) to develop coherent requirements and policies for designing zero trust enterprise infrastructure and procuring components.

The driver for Sections 2.1 and 3.1 is an initial attempt to form a neutral base of terms and concepts to describe ZTA. The abstract ZTA components and deployment models were developed to serve as basic terms and ways to think about ZTA. The goal is to provide a common way to view, model, and discuss ZTA solutions when developing enterprise requirements and performing market surveys. The above sections may prove to be incomplete as more experience is gained with ZTA in federal agencies, but they currently serve as a base for a common conceptual framework.

B.2.2 Perception that ZTA Conflicts with Existing Federal Cybersecurity Policies

There is a misconception that ZTA is a single framework with a set of solutions that are incompatible with the existing view of cybersecurity. Zero trust should instead be viewed as an evolution of current cybersecurity strategies as many of the concepts and ideas have been circulating for a long time. Federal agencies have been encouraged to take a more zero trust approach to cybersecurity through existing guidance (see Section 6). If an agency has a mature ID management system and robust CDM capabilities in place, it is on the road to a ZTA (see Section 7.3). This gap is based on a misconception of ZTA and how it has evolved from previous cybersecurity paradigms.

B.3 Systemic Gaps that Impact ZTA

These are the gaps that affect initial implementation and deployment of ZTA and continued operation/maturity. These gaps could slow the adoption of ZTA in agencies or result in fragmentation of the ZTA component industry. Systemic gaps are areas where open standards (produced either by a standards development organization [SDO] or industry consortium) can help.

B.3.3 Standardization of Interfaces Between Components

During the technology survey, it became apparent that no one vendor offers a single solution that will provide zero trust. Furthermore, it might not be desirable to use a single-vendor solution to

achieve zero trust and thereby risk vendor lock-in. This leads to interoperability within components not only at the time of purchase but also over time.

The spectrum of components within the wider enterprise is vast, with many products focusing on a single niche within zero trust and relying on other products to provide either data or some service to another component (e.g., integration of MFA for resource access). Vendors too often rely on proprietary APIs provided by partner companies rather than standardized, vendor-independent APIs to achieve this integration. The problem with this approach is that these APIs are proprietary and single-vendor controlled. The controlling vendor can change the API behavior, and integrators are required to update their products in response. This requires close partnerships between communities of vendors to ensure early notification of modifications within APIs, which may affect compatibility between products. This adds an additional burden on vendors and consumers: vendors need to expend resources to change their products, and consumers need to apply updates to multiple products when one vendor makes a change to its proprietary API. Additionally, vendors are required to implement and maintain wrappers for each partner component to allow maximum compatibility and interoperability. For example, many MFA product vendors are required to create a different wrapper for each cloud provider or identity management system to be usable in different kinds of client combinations.

On the customer side, this generates additional problems when developing requirements for purchasing products. There are no standards that purchasers can rely on to identify compatibility between products. Hence, it is very difficult to create a multiyear road map for moving into ZTA because it is impossible to identify a minimum set of compatibility requirements for components.

B.3.4 Emerging Standards that Address Overreliance on Proprietary APIs

As there is no single solution to developing a ZTA, there is no single set of tools or services for a zero trust enterprise. Thus, it is impossible to have a single protocol or framework that enables an enterprise to move to a ZTA. Currently, there is a wide variety of models and solutions seeking to become the leading authority of ZTA.

This indicates that there is an opportunity for a set of open, standardized protocols or frameworks to be developed to aid organizations in migrating to a ZTA. SDOs like the Internet Engineering Task Force (IETF) have specified protocols that may be useful in exchanging threat information (called XMPP-Grid [1]). The Cloud Security Alliance (CSA) has produced a framework for Software Defined Perimeter (SDP) [2] that may also be useful in ZTA. Efforts should be directed toward surveying the current state of ZTA-related frameworks or the protocols necessary for a useful ZTA and toward identifying places where work is needed to produce or improve these specifications.

B.4 Knowledge Gaps in ZTA and Future Areas of Research

The gaps listed here do not hinder an organization from adopting a ZTA for its enterprise. These are gray areas in knowledge about operational ZTA environments, and most arise from a lack of time and experience with mature zero trust deployments. These are areas of future work for researchers.

B.4.5 Attacker Response to ZTA

A properly implemented ZTA for an enterprise will improve the enterprise's cybersecurity posture over traditional network perimeter-based security. The tenets of ZTA aim to reduce the exposure of resources to attackers and minimize or prevent lateral movement within an enterprise should a host asset be compromised.

However, determined attackers will not sit idle but will instead change behavior in the face of ZTA. The open issue is how the attacks will change. One possibility is that attacks aimed at stealing credentials will be expanded to target MFA (e.g., phishing, social engineering). Another possibility is that in a hybrid ZTA/perimeter-based enterprise, attackers will focus on the business processes that have not had ZTA tenets applied (i.e., follow traditional network perimeter-based security)—in effect, targeting the low-hanging fruit in an attempt to gain some foothold in the ZTA business process.

As ZTA matures, more deployments are seen, and experience is gained, the effectiveness of ZTA in shrinking the attack surface of resources may become apparent. The metrics of success of ZTA over older cybersecurity strategies will also need to be developed.

B.4.6 User Experience in a ZTA Environment

There has not been a rigorous examination of how end users act in an enterprise that is using a ZTA. This is mainly due to the lack of large ZTA use cases available for analysis. There have, however, been studies on how users react to MFA and other security operations that are part of a ZTA enterprise, and this work could form the basis of predicting end user experience and behavior when using ZTA workflows in an enterprise.

One set of studies that can predict how ZTA affects end user experience is the work done on the use of MFA in enterprises and security fatigue. Security fatigue [3] is the phenomenon wherein end users are confronted with so many security policies and challenges that it begins to impact their productivity in a negative way. Other studies show that MFA may alter user behavior, but the overall change is mixed [4] [5]. Some users readily accept MFA if the process is streamlined and involves devices they are used to using or having with them (e.g., applications on a smartphone). However, some users resent having to use personally-owned devices for business processes or feel that they are being constantly monitored for possible violations of IT policies.

B.4.7 Resilience of ZTA to Enterprise and Network Disruption

The survey of the ZTA vendor ecosystem displayed the wide range of infrastructure that an enterprise deploying a ZTA would need to consider. As previously noted, there is no single provider of a full zero trust solution at this time. As a result, enterprises will purchase several different services and products, which can lead to a web of dependencies for components. If one vital component is disrupted or unreachable, there could be a cascade of failures that impact one or multiple business processes.

Most products and services surveyed relied on a cloud presence to provide robustness, but even cloud services have been known to become unreachable through either an attack or simple error. When this happens, key components used to make access decisions may be unreachable or may

not be able to communicate with other components. For example, PE and PA components located in a cloud may be reachable during a distributed denial-of-service (DDoS) attack but may not be able to reach all PEPs located with resources. Research is needed on discovering the possible choke points of ZTA deployment models and the impact on network operations when a ZTA component is unreachable or has limited reachability.

The continuity of operations (COOP) plans for an enterprise will likely need revision when adopting a ZTA. A ZTA makes many COOP factors easier as remote workers may have the same access to resources that they had on-premises. However, policies like MFA may also have a negative impact if users are not properly trained or lack experience. Users may forget or not have access to tokens and enterprise devices during an emergency, and that will impact the speed and effectiveness of enterprise business processes.

B.5 References

[1] Cam-Winget N (ed.), Appala S, Pope S, Saint-Andre P (2019) Using Extensible Messaging and Presence Protocol (XMPP) for Security Information Exchange. (Internet Engineering Task Force (IETF)), IETF Request for Comments (RFC) 8600. https://doi.org/10.17487/RFC8600

[2] Software Defined Perimeter Working Group "SDP Specification 1.0" Cloud Security Alliance. April 2014.

[3] Stanton B, Theofanos MF, Spickard Prettyman S, Furman S (2016) Security Fatigue. *IT Professional* 18(5):26-32. https://doi.org/10.1109/MITP.2016.84

[4] Strouble D, Shechtman GM, Alsop AS (2009) Productivity and Usability Effects of Using a Two-Factor Security System. *SAIS 2009 Proceedings* (AIS, Charleston, SC), p 37. Available at http://aisel.aisnet.org/sais2009/37

[5] Weidman J, Grossklags J (2017) I Like It but I Hate It: Employee Perceptions Towards an Institutional Transition to BYOD Second-Factor Authentication. *Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017)* (ACM, Orlando, FL), pp 212-224. https://doi.org/10.1145/3134600.3134629