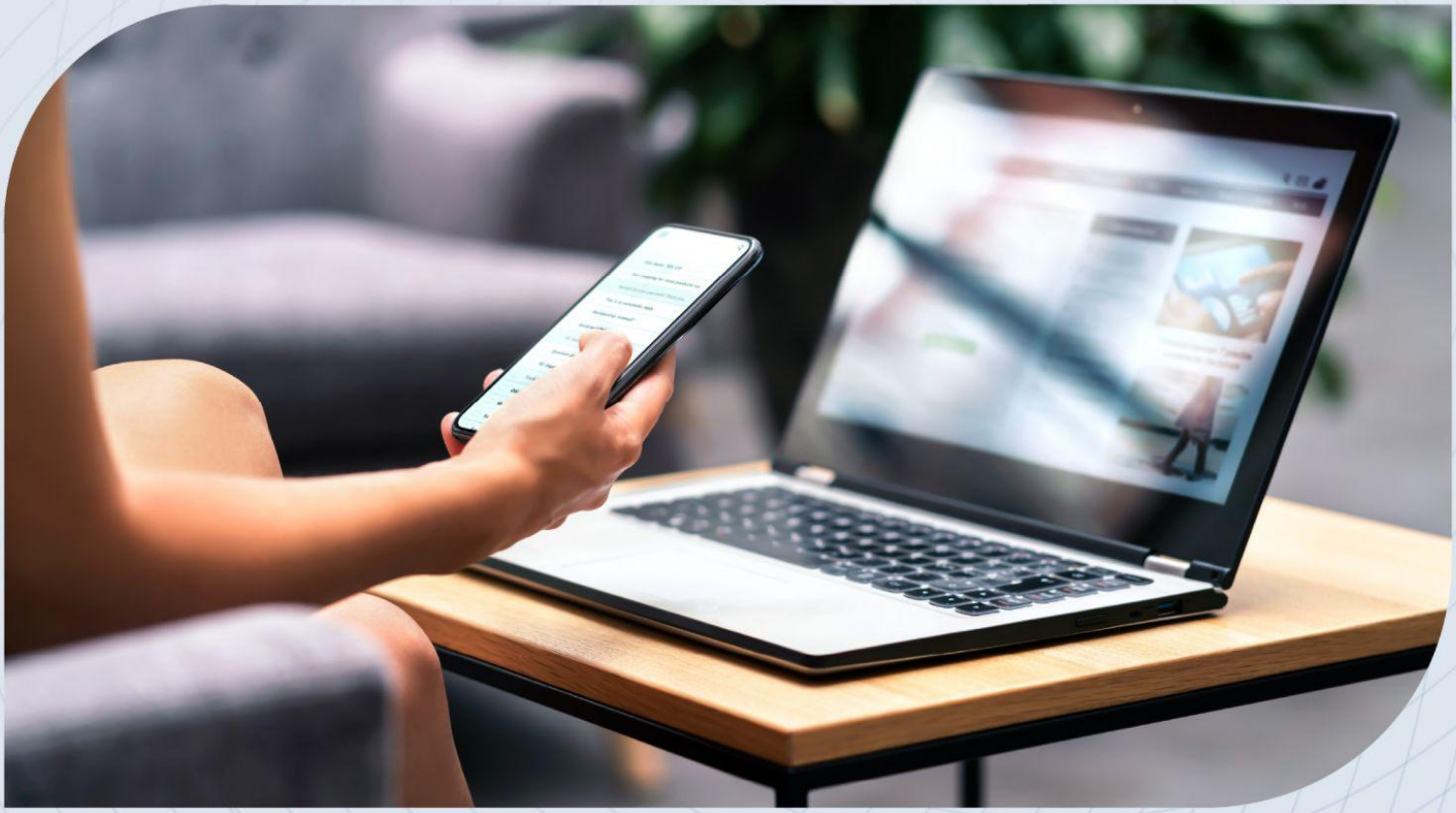
Zero Trust Principles and Guidance for IAM



Release Candidate

cloud security alliance®

This is a Release Candidate version and is subject to change.

Acknowledgements

The CSA Zero Trust Working Group

The scope of Zero Trust research and guidance necessarily includes cloud and on-premises environments along with mobile endpoints and is applicable to the Internet of Things (IoT) and operational technology (OT). The goals of the CSA Zero Trust (ZT) Working Group are to:

- Collaboratively develop and raise awareness of Zero Trust (ZT) best practices as a modern, necessary, and cloud-appropriate approach to information security (InfoSec).
- Provide thought leadership and educate the industry about the strengths and weaknesses of different ZT approaches so organizations can make informed decisions based on their specific needs and priorities.
- Take a deliberately product and vendor-neutral approach to architectures and implementation approaches for mature Zero Trust implementations.
- Take technically sound positions on Zero Trust and make defensible recommendations while remaining product and vendor-neutral.

Lead Author(s)

Alon Nachmany
Hani Raouda
Jonathan Flack
Kevin Dillaway
Paul Simmonds
Rohini Sulatycki
Shruti Kulkarni
Clement Betacorne
Irshad Javid
John Yeoh
Paul Simmonds

CSA Staff

Erik Johnson
Ryan Gifford
Stephen Lumpe

Reviewers

Anna Pasupathy

Table of Contents

Abstract

Identity, and the ability to consume information about that identity, as well as other Zero Trust (ZT) signals–additional attributes about an identity–is one of the key principles of Zero Trust architecture. A ZT approach aims to reduce the success of cyber-attacks and data breaches through risk-based access requirements. That is, by requiring authentication and authorization prior to granting access to resources (data and/or systems).

To meet this requirement, it is important to look at both existing and new identity, access management, and cloud solutions through a ZT lens.

ZT is a technology-agnostic guidance framework to bring controls closer to the asset being protected (the protect surface). From an identity, access management perspective, it offers increased capability of risk-based decision to grant access, instead of granting access based purely on the binary trust of a single access control method.

Target Audience

Primary: Technology managers for Zero Trust (ZT) implementation and architects
Secondary: CISO/ISO/Information Security, IAM vendors

Zero Trust Background and Drivers

Over the years, there have been various treatises that talk about trust as a human and social phenomenon, some of which used the term "zero trust." In 2001 the Open-Source Security Testing Methodology Manual (OSSTMM) began to address the issue of trust in information technology, and by its third (2007) edition, labeled 'trust' as a vulnerability, dedicating an entire chapter to the subject.

The concept of the Chewy Center (the smartie or M&M model) was introduced by Sun microsystems in the 1990s. In 2005-2007, the Jericho Forum (visioning paper and Jericho Forum® Commandments) and OpenGroup did some foundational work for Zero Trust on the failure of the traditional network perimeter security model and the need for de-perimeterization, which is the inspiration for the Open Group's Zero Trust Commandments.

Zero Trust Network (ZTN) concepts were developed by the US Department of Defense (DoD) in the early 2000s while defining Global Information Grid (GIG) Network Operations Black Core Network routing and addressing architecture, part of the DoD's Netcentric Service Strategy. Over time, this evolved into the ZTN Architecture (ZTNA) and Software-Defined Perimeter (SDP) framework that was embraced and subsequently further developed by the DoD, CSA, and NIST.

In 2010, after two years of research, John Kindervag of Forrester Research formally consolidated these concepts into the comprehensive area of practice we now know as Zero Trust. John's work was unique in that it formally identified the controls required to successfully implement these architectures, and provided an understandable method for implementing Zero Trust, including developing effective policies, leveraging the Kipling Method, and enabling expanded authorization controls, such as context-based access.

The DoD began to embrace Zero Trust in the 2019 timeframe after intelligence consultations with NSA concluded that the then-current methods were no longer effective, and the US needed to evolve its security strategy to better defend against increasingly sophisticated cyber attacks.

In August of 2020, NIST published SP 800-207 Zero Trust Architecture. In May of 2021, US President Biden issued Executive Order (EO) 14028, specifically mentioning Zero Trust security practices in his mandate to Federal Agencies to enhance cybersecurity, providing the first significant mandate for any government to adopt Zero Trust. While global interest and implementation of Zero Trust has increased in recent years, the US currently leads in Zero Trust adoption and the creation of associated guidance, largely as a result of such government mandates.

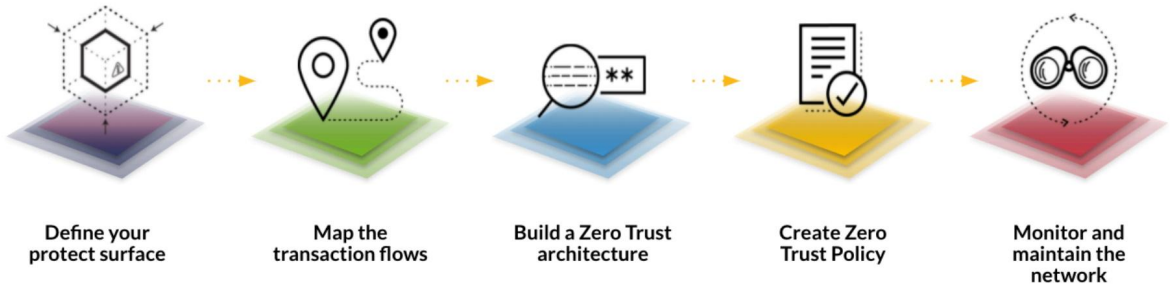
Related guidance, whether from NIST, DoD, CISA, or contributing experts at organizations like the Cloud Security Alliance, Forrester Research, or the UK NCSC, is ultimately based on the same fundamental

principles (as initially described in John Kindervag's foundational research), many of which are well established information security concepts (e.g. "least privilege", "deny all, permit by exception").
A key thing to understand about Zero Trust is that it is not a prescriptive architecture or a single product. Zero Trust is a strategy and a series of guiding principles that inform architectural and procurement decision. This enables organizations to design from the inside-out, based on their specific business requirements, assets, risks, and priorities.

ZT Implementation Methodology

National Security Telecommunications Advisory Committee (NSTAC) describes ZT implementation as a 5-step process. The five steps include:

- Defining protect surface
- Mapping the transaction flows
- Building a Zero Trust architecture
- Creating a Zero Trust Policy
- Monitoring and maintaining the network



MERRITT , R. (2022, June 7). What Is Zero Trust?
https://blogs.nvidia.com/blog/2022/06/07/what-is-zero-trust/

For more detailed information on the 5-step process, reference the NSTAC Report to the President on Zero Trust and Trusted Identity Management.

For ZT to be strategized, planned, and implemented, it is important to identify the organization's protect surfaces.

Scope

The scope of this paper includes looking at Identity and Access Management through the ZT lens in a technology-agnostic way and, as such, does not detail any engineering solutions.

The document explains the need to use identity attributes and other signals to drive the process of "authentication and authorization before granting access."

This paper refers generically to "entities", which refers to both persons and non-persons. From a perspective of Identity and Access Management, both entities have identity attributes and signals that provide an enhanced level of contextual awareness of risk.

Introduction

Authentication is the process by which one entity, such as a human, animal, object, device, network, application, database, process, service, and so on, can prove it is who it claims to be to another entity. NIST defines authentication as "verifying the identity of a user, process, or device, often as a prerequisite to allowing access to resources in an information system."

To enable authentication to withstand common authentication attacks like phishing, additional barriers are introduced to make it harder for the malicious actor to successfully compromise the flow. For example MFA strengthens authentication by introducing additional security barriers that make it harder for a hacker to successfully compromise an MFA-enabled authentication flow.

Authorization usually takes place after successful authentication and it enables the authenticated entity to access resources based on best practices like role-based access and the least privilege principle. NIST defines authorization as a right or a permission that is granted to a system entity to access a system resource.

Just like MFA strengthens authentication, Zero Trust strengthens authorization, by adding context awareness to the attributes that enable authorization to take place. The rest of this paper describes how Zero Trust achieves security and what it takes to implement Zero Trust in a scalable manner.

Identification of Entities and Attributes

Identity is a key element to delivering a Zero Trust security architecture, as it provides the attributes and signals for verifying and granting access to resources.

In a Zero Trust model, a request is not assumed to be trustworthy based on location, network or asset ownership but rather it is explicitly verified using multiple factors, such as the entity making the request, behavior, biometrics, cryptographic signature verification, location, and device health, operating system health, with each factor (ideally) understood to a known level of confidence.

Decision to allow access is made after the evaluation of the attributes and signals which should be context-aware and adaptive. Meaning that it can adjust the level of assurance required, based on the risk of the request and verification rarely being a one-time event, but a continuous process.
Ideally all access should follow the principle of least privilege.

By utilizing Zero Trust principles, organizations can reduce their attack surface, minimize the risk of breaches, and enable a more productive and flexible workforce.

One of the key principles of a ZT strategy is to be able to understand the context of requests and signals, enabling a better risk-based access decision to be made. Instead of relying on static credentials or roles, the system evaluates the context of each request based on dynamic attributes such as:

- User and groups they belong to (Who)
- Location (Where)
- Device (What)
- Time (When)
- Application type (How)
- Behavior and risk level of the user and the resource

Based on the context, the system can enforce risk-based granular policies and apply adaptive authorization mechanisms to ensure that only the right entities have access to the right resources at the right time and under the right conditions. This reduces the attack surfaces, chances of identity and credential theft, and can improve overall user experiences.

The challenge in any ZT architecture is to properly manage any repository for entities and attributes for which your organization is truly authoritative. It is also important to ensure that your organization (or systems they own and manage, such as cloud systems) are capable of consuming trusted attributes and signals with known levels of confidence from their authoritative sources.

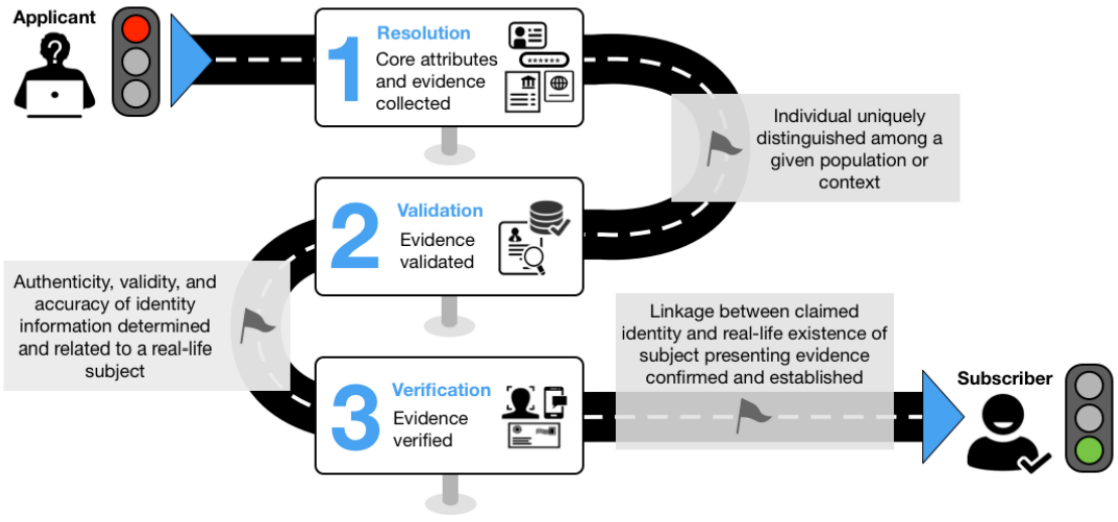
Identity Proofing and Validation:

Where an organization is authoritative for identity attributes and/or other signals used in a Zero Trust solution, robust processes used for the management and maintenance of those attributes and signals become critical. If a provably robust process is not used, then that organization risks not being trusted by any third party that needs to consume or rely on those attributes and signals.

Having a good level of assurance starts with a good proofing and validation process. NIST SP 800-63A defines the flow for identity proofing and enrollment to be a three-step process, but it's more often a five- or six-step process:

- Resolution where some attributes and evidence are collected
- Attestation or validation where the evidence collected is reviewed to determine that it is authentic, accurate, current, and unexpired
- Verification where a comparison between the evidence collected and the identity (entity) behind the future digital identity is done
- Digital identity provisioning where the digital identity is created into the source of truth (mostly IdP)
- Credential provisioning where the digital identity will be associated to one or more authenticators
- Digital identity deprovisioning where the digital identity is removed from the source of truth

One illustration (NIST SP 800-63A) of an authoritative source can be seen below:



Ref: NIST SP 800-63A

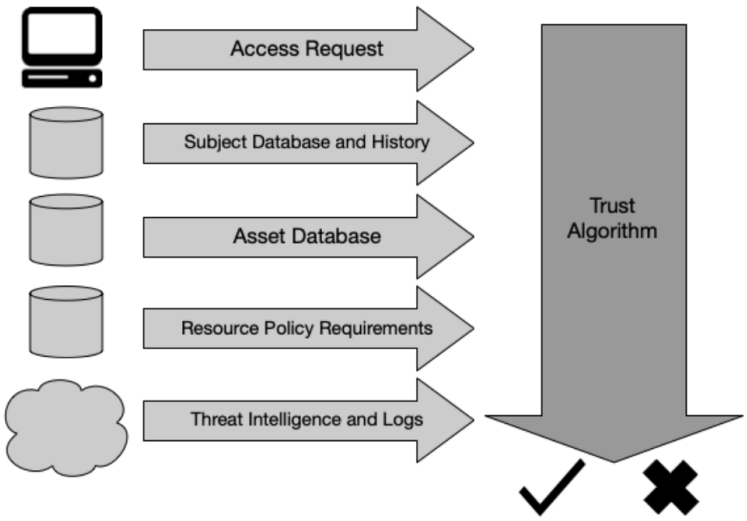
The control put in place during the process will be based on risks for the Data sources, Assets, Applications, and Services (DAAS) elements within an environment. This means a proofing and validation process should be available depending on the requirement of each DAAS component.

#	Function	Description	Steps in Phase
1	**Identity attestation**	Verify authenticity of identity documentation	• Verifiability can be manual, automatic, or a combination of both • • Define type of acceptable identification (e.g., driver license, passport, X509 digital certificate, token, etc.) • • Define the process, infrastructure, and tools needed to verify identification authenticity
2	**Identity provisioning**	Provision the validated identity into the source of truth (AD, IdP)	• Define the identity attributes needed to authenticate the users, in addition to user name and password (MFA, SSO, Passwordless, etc.) ○ • Define technology constraints over these selected attributes (password length, MFA technology, passwordless, etc.) • • Define a due-diligence process to ensure that the provisioned identity is not part of an internal or external blacklist (e.g., PKI revocation list)
3	**Credentials provisioning**	Provision users into common access solution	• Define how the solution will integrate with a rules engine that can exchange signals for authorization purposes • Securely integrate AD/IdP with solution • *Further information in Reference section below
4	**Identity deprovisioning**		• Triggered by Admin or code • Triggered by signals from AD or Identity attribute change

Raouda, Hani. (2023). Zero Trust identity and credential management . Retrieved June 10, 2023

Signals for Decision

The minimal signals as depicted by NIST SP 800-207 are shown in the figure below. Note that these signals are not required to all emerge from the same entity (e.g., vendor, provider, technology stack, etc.). However, each source's authenticity should be (cryptographically) verifiable and the signals it produces must be reliable, scalable, and tamper-proof and have an understandable level of confidence for use within the decision/risk process. Dependent on the risk-level, more signals may be required to offer more advanced contextual awareness, such as dynamic user behavior (e.g., keystroke patterns) or special purpose needs e.g., Deep packet inspection). Please note that the below diagram is for a trust algorithm. But the trust algorithm is the process with which the policy enforcement point grants or denies access. Thus, the output of the trust algorithm depends on the signals and the inputs it receives from the resources, like access request, asset database, and others as depicted in the figure below.



NIST. (2020, August). NIST SP 800-207.
https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf

Access requests from a subject to a resource should include:

- The resource requested
- Information about the requester including, but not limited to:
  - OS version
  - Software used
  - Patch level
  - Duration of access (just-in-time)

Subject database and history show "who" is requesting access to a resource. This is a collection of subject attributes and assigned privileges, e.g.:

- User attributes (e.g., account ID)
- Results of authentication checks performed by PEPs
- Roles and permissions

Asset database contains the known status of each enterprise-managed asset. This is compared to the observable status of the asset making the request and can include:

- OS version
- Software present
- Asset integrity
- Location (network location and geolocation)
- Patch level

Depending on the asset state compared with this database, access to assets might be restricted or denied.

Resource policy requirements define the minimal requirements for access to the resource. This set of policies complements the user ID and attributes database. Requirements may include authentication assurance levels such as:

- Network location (e.g., deny access from overseas IP addresses)
- TLS 1.2 and above
- Data sensitivity
- Requests for asset configuration

These requirements should be developed by both the data custodian (i.e., those responsible for the data), data owner, and those responsible for the business processes that utilize the data (i.e., those responsible for the mission).

Threat intelligence and logs provide information feeds about general threats and active malware from various sources. This can include indicators of compromise seen on the device, such as queries for possible malware command and control nodes, and communication with command-and-control sites.

- Threat intelligence feeds can be external services or internal scans
- Discoveries can include attack signatures and mitigations

This component will most likely be under the control of a service rather than the enterprise.

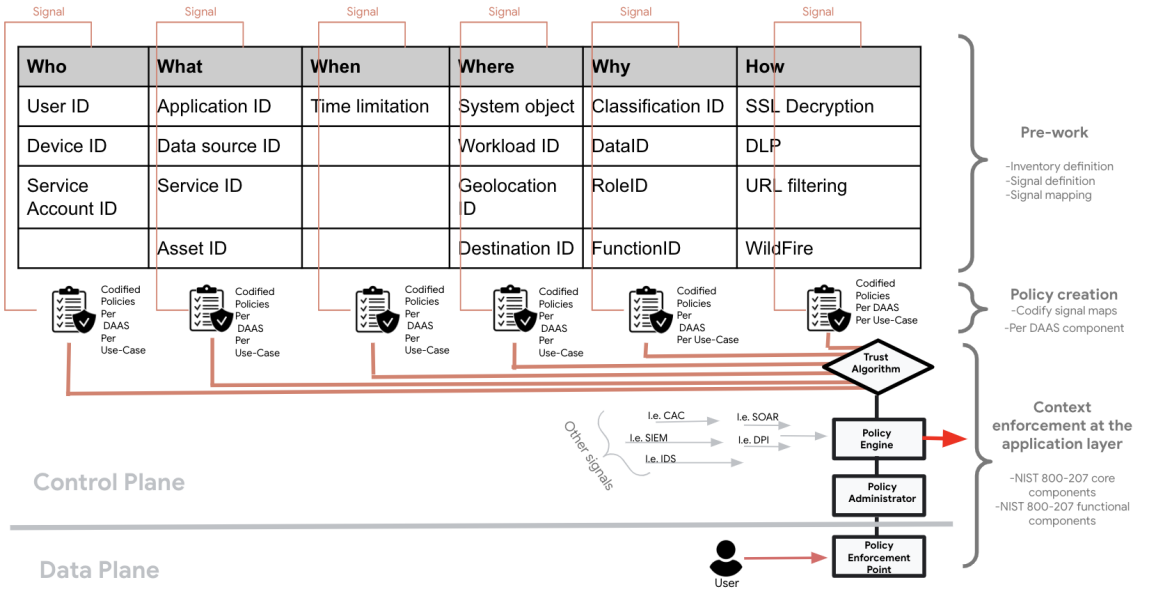
Decision rules: The **Five Ws,** sometimes referred to as **Five Ws and How**, **5W1H**, or **Six Ws**, or the **Kipling Method**, can be used to obtain at least 6 dimensions of context that can be used during the authorization process through a rule's engine.

NIST SP 800-207 describes the core Zero Trust components (Rules Engine → Policy Engine, Policy Administrator, Policy Enforcement Point, Trust Algorithm) that can be employed to deliver this context-based authorization scheme.

Every protect surface (e.g., a single DAAS component that is to be protected with Zero Trust) will have incoming and outgoing signals into it, and from it, respectively. These signals form the ingress and egress traffic. That traffic could be going to the public, to employees, or to another protect surface, such as a database, a server, or an API.

Every outgoing and incoming data packet from/into a protect surface must be verifiable by the rules engine, according to a policy that codifies the right combination of those signals (Who, What, When, Where, Why, How). The rules engine can receive other signals related to user behavior, device posture, deep packet inspection results, DLP results, and so on.

The rules engine that is part of the Zero Trust authorization process will leverage its algorithm to grant context-aware authorization at potentially all layers of the OSI Model.



Raouda, H. (2023). Zero Trust context aware authorization . Retrieved June 10, 2023,.

The goal of documenting such signals and their corresponding policy code per protect surface, per use case, is to allow security engineers to easily design, implement, test, and version-manage the ZT contextual awareness for a given protect surface.

The image above is just to provide some clarity on how signal definition can be easily followed by codifying these constraints into policy as code, which can then be easily consumed by the rules engine in place.

Authorization Based on Policy

In ZT, we look for risk-based access controls which incorporate concepts from old school Role Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) but is much more than these concepts, which were generally reliant on a single source of truth managed by the organization.

However, when designing a ZT architecture, it is crucial to look at existing and new identity and access management solutions from a ZT perspective.

Policies provide the first point of contact for authorization. Policies are based on transaction flows between the requestor and resources. When implemented on a ZT solution, granular policies should be created. For example, entity A is allowed to access resource B from a defined device between 9:00 a.m. and 5:00 p.m. standard time.

Policies define explicit authorizations in a ZT environment and are performed as close to the DAAS elements as possible which require controls and policies that they can apply, with these policies leveraging signals received from a number of sources, where one of them may be an Identity Provider (IdP).

Dealing with Failed Policy Decision

When the conditions to grant access are not met, the following typical solutions may be employed (this is a non-exhaustive list):

- The entity can be black-holed
- The entity is informed access has been denied
- The entity can be put into a queue

- The entity can be suspended for a period of time
- Step-up authentication can be required
- Notifications can be sent to the policy administrator to verify the need for policy refinement
- Relevant data can be sent to the SIEM for further analysis which can then be used to refine policies or alert for possible attacks
- The entity can be sent to a honeypot

By implementing these additional workflows, organizations can effectively manage failed policy decisions, maintain a secure environment in line with the principles of Zero Trust, and adapt their security measures based on real-world scenarios and evolving threats.

Business Value

Note: This section does not cover the generic business value of Zero Trust. Refer to *Communicating the Business Value of Zero Trust* for more details. The business value of identity, in relation to Zero Trust, is listed below.

In a Zero Trust environment, controlling access to data and/or system (resource entitlement) based on identity, can provide several business benefits:

- **Improved security**: By controlling access to resources based on identity, organizations can reduce the risk of unauthorized access to sensitive data and resources, minimize lateral movement and therefore minimize the potential for system compromise and data breaches.
- **Improved compliance**: Many regulations and standards, such as HIPAA and GDPR, require organizations to implement strong access controls to protect sensitive resources. By using a set of granular entitlement rules an organization can clearly demonstrate compliance with these requirements.
- **Reduced friction**: By implementing identity-based entitlement control, entities can enjoy a more frictionless and secure access experience allowing the access to the resources needed without having to repeatedly enter credentials or navigate multiple control layers (barriers).
- **Increased agility**: By correctly aligning business requirements to identity-based entitlement controls across multiple entities organizations can more quickly and easily adapt to changes in the business environment, such as new employees joining the organization or new applications being deployed, or if implemented correctly, allowing entities from outside the organization to (simply and directly) access permitted systems and data without the need to create “dummy” users or implement special gateways or other access control layers.
- **Increased productivity**: By correctly aligning business requirements to identity-based entitlement controls across multiple entities, users should then automatically have the access

they require to be productive (and nothing more); without the need to get permission from IT to access resources.

- **Reduced costs**: By reducing the number of unauthorized access attempts, organizations can reduce the costs associated with investigating potential and actual security incidents, data breaches and ransomware attacks.

Conclusion

In the past, access to computer systems was based on implicit trust; the user was trusted because they gave the correct password, and the computer they were using was trusted because they were on a known network (usually the organization's Intranet). The entire organization’s environment was architected around this principle because it made sense when it was initially created, and access to systems outside of the closed environment were minimal.

The problem is that over the last two decades, there has been a migration to computing outside of that monolithic environment,; driven by the adoption of more cloud services and the increasing need for organizations to collaborate with entities that they do not own and are not part of their sphere of control.

Thus the traditional strong perimeter model has become harder to configure, enforce, afford, and maintain.

The Zero Trust model starts from the point of do not trust, but rather verify access to data and systems based on risk. This extends beyond human entities to encompasses devices, organizations, code, agents, and service-based identities.

With Zero Trust, the paradigm shifts from a binary trust, to an adaptive authentication and authorization model. Access to systems and data is granted based on a set of identity attributes offered by that entity. Together with other intelligence (signals) that combine in real-time to provide a confidence level that the access meets (or exceeds) the risk-level set for access to be granted.

Furthermore, access is continually reevaluated based on a frequency determined by the sensitivity of application, solution, service, and/or device. The result is an environment that has the potential to provide increased security as well as provide the organization with a more flexible (frictionless) IT environment.

Once processes are developed and the model is implemented, it can provide more telemetry and insight into what is happening in the environment and allow organizations to be more responsive to potential issues. It also can be implemented in a phased approach to make it achievable.
In this paper, we have explained Identity and Access Management in the context of Zero Trust and we hope this will help you get started on your Zero Trust journey.

References

Signal - Access Context Signals provide organization administrators the inputs required to satisfy fine-grained, attribute based access controls which validate all authorized access to the DAAS elements within a protect surface; answering the questions What and Where and When, How and Why, and Who.

Many foundation resources are capable of providing effective access context signals. The Identity Provider (IdP) provides a strong signal asserting identity, as well as adherence to organization authentication policy (multifactor, CAC, etc.), answering the 'Who?' question. The request can tell us 'What' and 'Where' a resource is being accessed.

From these signals, we can easily construct a policy which can enforce access.

NSTAC - Section 3.3.3: NSTAC REPORT

Definition of protect surface: NSTAC defines the protect surface as the area the ZT policies protect. Each protect surface should ideally contain a single data, applications, assets, and services (DAAS) element, and in turn, each ZT environment will have multiple protect surfaces.

https://media.defense.gov/2023/Mar/14/2003178390/-1/-1/0/CSI_ZERO%20TRUST%20USER%20PILLAR.PDF

https://docs.google.com/document/d/1yMH8vcT0ROwtXG4n8uYPibNLspjAul3Opiut0xnyuKg/edit?usp=sharing

One example of a flow of credential provisioning and integrating the IdP with a solution is for each identity that requires credentials, is as below:

- Issue any number of public key pair(s) to be used by the provisioned identity for digital signing
- Send public keys of each identity to source of truth AD/IdP as one of the attributes
- Issue Identity symmetric key(s) to be used by identity for encryption
- Store private keys and symmetric keys within a device TPM
- Issue Identity Public X.509 certificate or private certificate
- Sign certificates with the CA of the organization
- Store identities and their corresponding public keys on an immutable ledger (can be private and/or managed)
- Create mechanisms for other solutions to securely obtain the public keys and certificates for a given identity

Foundational References

NSTAC REPORT

NIST SP 800-63: This document outline the importance a doing a risk assessment specifically on digital identity considering enrollment, identity proofing, authentication and federation

NIST SP 800-63: ZT maturity through the user pillar (NSA just published Advancing Zero Trust Maturity Throughout the User Pillar: CSI_ZERO TRUST USER PILLAR.PDF (defense.gov))

Advancing Zero Trust Maturity Throughout the User Pillar: This document break down the identity pillar into 5 categories (Identity management, Credential management, Access management, Federation and Governance) and define for 3 categories (Identity management, Credential management and Access management) a maturity level from preparation to Advanced ZT maturity
Kipling Method: [ref: https://en.wikipedia.org/wiki/Five_Ws]

https://www.paloaltonetworks.com/blog/2019/05/network-layers-not-created-equal/

https://federalnewsnetwork.com/wp-content/uploads/2020/01/simplify-zero-trust-implementation-with-a-five-step-methodology.pdf

https://cloudsecurityalliance.org/cloud-security-glossary/