Zero Trust

Going Beyond the Perimeter

Zero Trust

Going Beyond the Perimeter

AUTHORS

J. Wolfgang Goerlich
Wendy Nather
Thu Pham

0.0

Why Zero Trust?

The invisible line that we draw between what belongs to the enterprise and what doesn't — servers, desktops, networks, applications, and logins — traditionally depends on firewalls and endpoint-resident security software to protect that boundary, but the headlines are full of examples where that simply wasn't enough. People have certainly been promoting the perimeter's demise for years now: the **Jericho Forum** was created to tackle "de-perimeterization" as early as 2003. The idea really picked up steam as the cloud became more accepted as a common place to store and process data. John Kindervag at Forrester Research coined the term "zero trust" around 2009 to propose a specific framework. Google described in detail how they implemented the principle internally, and dubbed it "BeyondCorp." It's now within practical reach for many more organizations, with a concrete example to consider implementing.

The idea of getting rid of the perimeter is generally too scary for enterprises to contemplate, especially if they've only recently solidified one.

So let's not think of it as getting rid of the perimeter, but rather as tightening security on the inside so that the network perimeter isn't the only thing keeping the attacker at bay.

Traditional Approach

The classic approach to securing corporate information resources assumed several things:

1. Every endpoint being used to access resources was owned, issued and managed by the enterprise.

2. All users, devices and applications were in fixed and predictable locations, usually on a corporate network behind a firewall.

3. One method of verification at the point of initial access was sufficient.

4. Corporate-managed systems with the same classification could all inherently trust one another.

Over the years, we've come to acknowledge that these assumptions no longer hold true, thanks to mobility, BYOD (bring your own device), cloud, and increased collaboration among partners. The consumerization of IT has prompted users both to demand a more customized environment and to insist on using their personal devices without corporate management. Attackers that make it past one verification point (such as a firewall or a user login) can exploit inherent trust and move laterally within a network, application or environment to target sensitive data. An insider that starts within a trusted zone can escalate privileges. **We can no longer assume that "internal" entities are trustworthy, that they can be directly managed to reduce security risk, or that checking them one time is enough.**

Move Toward Zero Trust

Kindervag defined the guiding principle for "zero trust" as "never trust, always verify." In other words, assume that every part of your network is potentially hostile, as if it were directly on the internet, and treat access requests accordingly. Threats that manage to bypass the firewall (through compromised user credentials or a vulnerable web-facing application, for example), or that start within the internal "trusted" network, should be stopped by additional security controls that prevent lateral movement and thereby minimize the impact of a breach.

Instead of thinking of the perimeter as one type of access control around the "edge" of the network, **think of the perimeter as any place where you make an access control decision**. This could still be at the firewall or switch, but it could be at other layers as well: the difference between logging in to a third-party SaaS application with a personal ID and logging in with a corporate ID dictates which security decisions apply, and who makes them. Where an application tries to access a database, that's a perimeter. When a user elevates privilege to perform a sensitive operation, that's also a perimeter. The zero-trust model of security prompts you to question your assumptions of trust every time there's an access event.

The Zero Trust Approach

A zero-trust model is built upon the following fundamentals:

- **Visibility informs policy.** Provide as much intelligence and insight as possible to the people administering the technology, in order to produce informed policies.
- **Trust is neither binary nor permanent.** Continually reassess the posture of users, devices, and applications and adjust your trust accordingly. Be prepared to respond to events that raise the risk level by containing newly discovered threats and vulnerabilities.
- **Ownership is not a control.** Validate and extend trust to devices, applications, and networks that you don't own or manage, from BYOD and IoT (Internet of Things) devices to SaaS and public cloud.
- **The perimeter is any place where you make an access control decision.** Choose the layers and process points that work for your environment, whether it's at the network layer, the application layer, at the point of identity verification, or during a transaction workflow.
- **Access decisions are based on re-establishing trust every time.** Membership within a group, an application service within a tier, or a device connected to a network location, are not enough on their own to authorize activity.
- **Containment.** Combine least privilege and segmentation with response capabilities to monitor for threat activity and limit its spread by default.

In addition to questioning all assumptions of trust, your implementation should ideally include these characteristics:

- **Transparency.** Security is as invisible as possible to people using the technology[1].
- **Zero-touch for zero trust.** Minimize the administrative effort through rationalization, automation, orchestration, and integration.

BUSINESS OUTCOMES

With the zero-trust model, you gain **better visibility** across your users, devices, containers, networks, and applications, because you're verifying their security states with every access request.

You can reduce your organization's attack surface by segmenting resources and only granting those permissions and traffic that are strictly needed. And by using more authentication factors, adding encryption, and marking known and trusted devices, you can **make it harder for attackers** to collect what they need (user credentials, network access, and the ability to move laterally).

Finally, your users can get a consistent and more productive security experience regardless of where they're located, what endpoints they're using, or whether their applications are on-premises or in the cloud.

[1] Some experts have also described this as "translucency:" it should be just visible enough so that if necessary, users can reassure themselves that it's there.

Introducing the Three Pillars of Zero Trust

Security is not a one-size-fits-all proposition, even within the same enterprise environment. For example, continuous authentication is a great idea, until it conflicts with users having a low-friction workflow: if they have to authenticate with multiple factors too often, they'll resent it (and try to evade the controls that require it). Software itself, on the other hand, doesn't mind frequent authentication, so workloads that communicate with one another can support those interactions. IoT devices, such as medical or manufacturing equipment, can have both safety and availability implications that affect how they are connected to a network. We introduce three pillars of zero-trust security to outline the differences:

01

Zero Trust for the Workforce

People such as employees, contractors, partners and vendors accessing work applications using their personal or corporate-managed devices. This pillar ensures only the right users and secure devices can access applications, regardless of location.

02

Zero Trust for Workloads

Applications running in the cloud, in data centers, and other virtualized environments that interact with one another. This pillar focuses on secure access when an API, a microservice or a container is accessing a database within an application.

03

Zero Trust for the Workplace

This pillar focuses on secure access for any and all devices (including IoT) that connect to enterprise networks, such as user endpoints, physical and virtual servers, printers, cameras, HVAC systems, kiosks, infusion pumps, industrial control systems, and more

In the following sections, we break down each pillar by the risks addressed, options for implementation, and proposed maturity levels.

	WHO OR WHAT	TRUST GETS VERIFIED WHEN	FROM
WORKFORCE	People & Their Devices	Accessing Applications	Anywhere
WORKLOAD	Apps, Services, Microservices	Communicating with Other Systems	On-Premises, Hybrid Cloud, Public Cloud
WORKPLACE	IT Endpoints & Servers, Internet of Things (IoT) Devices, Industrial Control Systems(ICS)	Accessing the Network	On-Premises, Hybrid Cloud, Public Cloud

1.0
Zero Trust for the Workforce

RISKS ADDRESSED

Zero Trust for the Workforce addresses several important risks for the enterprise:

- Primary account credentials (username and password) are often stolen through phishing attacks or compromised third parties, and re-used by attackers from remote locations, including botnets. According to the ***2019 Verizon Data Breach Investigations Report***, nearly one-third of breaches involved compromised credentials, showing that passwords are an effective way to get past traditional perimeter defenses and get access to applications, undetected.

- An attack that can bypass the firewall, or that starts on the internal network, can spread out to compromise critical systems and steal sensitive data. **And let's face it: a sufficiently successful outsider looks exactly like an insider.** An external attacker will use the same means to get in that work for the legitimate user, so you have to make sure to limit what everyone can do.

- Another risk is that the attacker exploits the gaps between different policies or enforcement that apply to the same asset. If the same confidential data is available in two different systems using different types of authentication, the attacker will go after the one that's easier to reach – either because it trusts something else that you can leverage, or because that one authentication method has a flaw in it. When an application or system is protected with different controls dependent on whether the user is "inside the perimeter" or not, an attacker can compromise the looser set of controls.

- External cloud-based applications and mobile users can face attacks that are outside of the enterprise perimeter protections.

- Users can make the organization vulnerable by using unmanaged and unpatched devices to connect to critical systems and data. These weaknesses can lead to ransomware and other kinds of malware attacks, as well as unauthorized access.

OVERVIEW

The Zero Trust for Workforce implementation rests on the combination of validated users using validated endpoint devices. This combination is further locked down with end-to-end encryption between these devices and the resources they access.

Finally, users are allowed only the bare minimum access needed for their roles (which is also known as "least privilege"). As long as the user is authenticated with the right number of factors, and is using an endpoint that has been enrolled and inspected for security vulnerabilities, they can access exactly those resources that they're allowed to by a centralized proxy.

WORKFORCE MATURITY MODEL

STAGE 1 ESTABLISH USER TRUST

Ensure you have the right mechanisms and processes to ensure only authorized users are attempting to access your resources. This can be achieved in a number of ways, but multi-factor authentication (MFA) is a commonly used technology.

STAGE 2 DEVICE AND ACTIVITY VISIBILITY

Which endpoint or device is being used with every access request? What is its current security state, and where is the request originating? This is a key stage for detecting account takeover attempts and other risks.

STAGE 3 TRUSTWORTHY DEVICE

Whether it's corporate-owned or not, whether it's managed or unmanaged, an organization can mark as trusted the devices it has registered and expects to see associated with that particular user.

STAGE 4 ADAPTIVE POLICIES

Implement requirements for access based on the sensitivity of the resources and the known security state to manage risk levels appropriately. These policies can range from allowing only corporate-managed devices to requiring certain versions of patched software, encryption, or step-up authentication based on user behavior.

STAGE 5 ZERO TRUST FOR THE WORKFORCE

At this point, all applications and systems are covered by the previously listed stages; monitoring and response to risk events are going on continuously; and the users get a consistent single sign-on experience.

2.0
Zero Trust for
Workloads

RISKS ADDRESSED

Zero Trust for Workloads addresses several important risks for the enterprise:

- An attacker exploiting application vulnerabilities can move laterally to compromise critical systems.
- An attacker obtaining sensitive data and exfiltrating the data out of the network.
- Disparate controls between internal applications and external cloud-based applications creating blind spots for defenders.
- Developers making the organization vulnerable through the code and configuration of web applications.
- **Fifty-four percent of web app vulnerabilities have a public exploit available**, meaning if servers and applications aren't patched, they're left open to known flaws that can be exploited by an attacker to get access to your systems.

OVERVIEW

Enterprise systems tend to be organic: they grow in functionality and add connections and dependencies in response to business needs. In order to facilitate this growth, system designers and developers can sometimes trend towards the most permissive and flexible security configurations. This creates excessive trust which attackers may exploit and move laterally to access sensitive resources.

The textbook answer to this challenge is network segmentation. Consider a generic three-tier web application: Presentation tier, Application tier, and Data tier. These tiers can be segmented into different networks, with specific access controls restricting how the tiers communicate. Even in this simple example, services on the Application tier will be trusted to communicate to other services on the same tier, doing little to address the risk of lateral movement within the tiers. The problems of excessive trust increase with the complexity and the number of applications. Moreover, over time an organization can lose track of which workloads are critical and what other resources need to communicate with them, making it harder to lock them down.

There are two philosophical ways to address the concerns. We can assume the network is untrusted and move the trust decision up into the application stack. The benefit of this philosophy is we put the control into the application. The disadvantage is applications must be developed with this zero-trust approach in place. Developers don't always document how an application should communicate across its own workload, much less with external resources; this makes it more difficult for network and security operations teams to know how to balance least privilege and application availability.

Alternatively, we can reduce the trust in the network by tightening the communications to only what is needed for the application. This philosophy works well for existing applications, including legacy applications, and can be a way to bootstrap an existing ecosystem into the zero-trust model. The disadvantage here is that we are reliant on the network for security and, should that security be compromised, the application services will be unaware of the reduced security posture.

These philosophies are not mutually exclusive, and in fact they may both be useful to balance each other in an environment where redundant controls are needed for high assurance.

For bootstrapping existing environments into zero-trust models, our networks must evaluate trust and make access control decisions at the point of network communication. This is no simple feat considering our application services often spread across cloud service providers, data centers, and other heterogeneous virtualized environments. We need to define an application ecosystem to contain only the application's dependencies: services, processes, and network communications. We can then apply access control using a whitelist or default-deny, such that only what the application requires is permitted regardless of the network or environment. **We define trust by the application's unique requirements, not by the network location.**

Achieving this micro-segmentation requires three technologies which have, up until recently, been out of reach.

- **Deep and pervasive visibility into network communications.** Distributed network sensors instead of traditional centralized monitoring (SPAN/TAP or NetFlow) have made such visibility possible at scale.

- **Accurate and real-time application modeling.** Big data analytics techniques have significantly reduced the manual effort in documenting applications, thus enabling up-to-the-moment understanding of traffic patterns and dependencies.

- **The ability to apply policy to multiple devices across multiple environments.** A high-level policy engine that manages the ongoing sprawl of access control devices across multi-cloud environments simplifies the steps required to act on the application visibility and analytics.

Combined, visibility, analytics, and policies reduce the excessive trust in the application ecosystems.

But what happens when even this trust is abused? Consider, for example, the risk an enterprise faces from the administrators and other privileged users, who tend to have elevated access at scale. Any intruder who compromises developer or administrator credentials could potentially get access without security operations noticing. Staffing security operations with people to inspect the individual workloads and connections doesn't scale. To achieve Zero Trust for Workloads, unsupervised machine-learning techniques and behavioral analysis are employed to monitor for signs of malicious activity. When identified, the network can revoke trust by quarantining servers and blocking communications.

When the velocity of change exceeds the capacity of people, the move to automation becomes inevitable. This is the state segmentation efforts find themselves in today. Adopting a zero-trust mindset enables system designers and developers to come at the problem in new ways. With better visibility, quicker analytics, and a deeper understanding of application communication, Zero Trust for Workloads redefines the perimeter around expected behavior. Malicious activity, from initial compromise to lateral movement to data exfiltration, then becomes apparent and preventable.

WORKLOADS MATURITY MODEL

STAGE 1 ESTABLISH WORKLOAD TRUST

Discover the application ecosystem and environments with mission critical workloads. This stage establishes the scope for the zero-trust initiative.

STAGE 2 WORKLOAD VISIBILITY

Gain visibility into the devices, processes, packets, network flows, and workload communications within the application environments. This effort is scoped to the application ecosystem, and visibility is crucial for gaining insights into the workloads (such as unpatched software and configuration states).

STAGE 3 MAP APPLICATION DEPENDENCIES

Analyze the network communications and data flows to model applications, categorize application tiers, and identify application dependencies. This is performed over a period of time to capture infrequent activities, such as monthly jobs or one per quarter accounting processes. The more accurate the application mapping, the more accurate the resulting policies will be.

STAGE 4 POLICIES AND MICRO-SEGMENTATION

Develop policy to minimize the trust within the application ecosystem, simulate and validate the policy, and deploy the policy consistently across all environments, taking into account identity and contextual information from the Workforce and Workplace pillars where appropriate. Micro-segmentation takes a traffic whitelisting approach, otherwise known as default-deny, to move the access perimeter to just what is needed for the workload.

STAGE 5 ZERO TRUST FOR WORKLOADS

Mature zero-trust organizations demonstrate continuous improvement and continued monitoring of the environments. Change is the only constant – change in the application, change in the organization, change in the attacks – and zero trust requires evolving the policies as the ecosystem evolves.

3.0
Zero Trust for the Workplace

RISKS ADDRESSED

Zero Trust for the Workplace addresses several important risks for the enterprise:

- An attacker exploiting endpoint, server, or facilities equipment vulnerabilities to gain a foothold in the network and move laterally to compromise critical systems.
- An attacker disrupting operations through attacks on networked business infrastructure.
- Weaknesses in IoT or Operational Technology (OT).
- Sixty percent of businesses have had security incidents stemming from network printers, **according to Quocirca**.
- There has been a **300% rise in new IoT malware variants from 2017 to 2018**, according to Kaspersky.

OVERVIEW

The modern workplace is enabled by the campus, data center, WAN, branch and cloud network. Trust is extended to any user, device and application, linked wired or wirelessly, to connect to other users, devices, applications and other parts of the workplace. The workplace encompasses end-user devices, IT servers and printers, Industrial Control Systems (ICS), and IoT devices. Zero Trust for the Workplace is enforcing trust when any kinds of devices are authenticating and communicating on the enterprise networks.

There is, however, a very real difference between devices used by the workforce and the equipment in our workplaces. The idea of enforcing trust on access decisions for end-user applications doesn't translate for equipment such as printers, manufacturing controllers, HVAC, and badge readers. In order to cover all business-related systems, we need to move lower in the stack to the network.

The rapid growth of devices on our networks have strained our ability to manage devices, patch devices, and protect against rogue devices. IoT gets much of the attention due to the explosion in network-enabled devices in recent years. IoT is often built on consumer-grade platforms, lacks enterprise-level security controls, and may not be patchable. The result is we have more of these devices, these devices have comparably more vulnerabilities per unit, and IoT is comparably more difficult to secure. While IoT is in the spotlight, we cannot overlook traditional business equipment such as printers, videoconferencing, security cameras, and VoIP telephony, which continue to be a viable avenue for criminals to compromise enterprises. Then, we also have medical equipment and OT to consider. These are often on platforms security teams cannot patch or secure, due to a number of operational, functional, and technical factors. Broadly speaking, a Zero Trust for the Workplace strategy must address authenticating, authorizing, segmenting, and monitoring trust across all equipment.

Zero trust assumes the network is inherently insecure. We need to protect the network from the users, devices and applications connected to it, and vice versa. In a zero-trust network, any exploitable device has to be shielded or segmented to reduce the likelihood of a criminal finding and exploiting the device. Moreover, in a zero-trust network, the remaining devices have to be protected from other compromised and exploited devices. These protections go hand in hand. Both require a known inventory of the entities using the network, and visibility into the security posture of the devices.

The access control decision occurs when equipment attempts to connect to the network. Traditionally, network engineers accomplished this with fixed attributes such as a combination of network switch location or IP address. In this model, we trust equipment without knowing whether the equipment is vulnerable or exploited. The traditional trust is also based on easily spoofable attributes. When moving to zero trust, the decision must be made on a number of factors, including identity and behavior, and it must be verified regularly based on device behavior and any changing factors. In particular, the organization must be able to respond to newly discovered threats and vulnerabilities by limiting the original network access or cutting it off altogether.

Network Access Control (NAC) forms the foundation of a zero-trust implementation. The equipment must authenticate to the network before it is trusted to connect and communicate. The ideal is software-defined access control built with 802.1X and certificate-based authentication. Windows-based devices can take advantage of Active Directory and Windows Management Instrumentation (WMI) to authenticate to the network. If these methods are not available, we can use MAC Authentication Bypass (MAB). MAB is spoofable; however, it may be the only option for older equipment which does not support newer methods, or equipment which we cannot configure to support newer methods.

The next level of a zero-trust network is group-based segmentation. We authenticate network connections. When making the access decision, the network identifies the equipment as belonging to one or more roles, and one or more groups. These roles are irrespective of IP addressing or physical location. In fact, in most complex enterprises, these roles include multiple subnets and multiple buildings. We then define segmentation policies based on which groups of entities can talk to which network resources, including the internet. Based on the behavior of the equipment, we can ascertain trust, and further restrict access to it when there is cause for concern. We can continue to reduce the assumed trust and strengthen security in the network through continuous monitoring of communications and continuous improvement of policy sets.

The multiplication of employee-driven devices has led to a corresponding increase of devices within our enterprise networks. From IoT to printers, from OT to medical devices, more equipment than ever is powering our organizations. Consequently, the attack surface from equipment is larger than ever. A Zero Trust for the Workplace strategy enables security operations and network engineers to have better visibility into all hosts and communicatons, provide tighter restrictions on network communications, and implement adaptive policies based on trust. We can then reduce the risk of malicious activity exploiting these devices, and respond quicker to any suspicious traffic.

WORKPLACE MATURITY MODEL

STAGE 1 ESTABLISH WORKPLACE TRUST

Discover workplace systems, their users and applications, including IoT and OT, and determine their function within the organization and their operation on the network. Define the scope for the zero-trust initiative.

STAGE 2 NETWORK VISIBILITY

Gain visibility into the user, device and application communications and network flows within the workplace environment. Understand and document the in-scope network capabilities and requirements.

STAGE 3 NETWORK ACCESS CONTROL

Configure and enforce network authentication and authorization for the in-scope users (where present), devices, and the applications. Prevent any unauthenticated (and therefore untrusted) entities from connecting to the in-scope network.

STAGE 4 SEGMENTATION POLICIES

Define group-based network policies which enable only those network connections and communications required for business operations.

STAGE 5 ZERO TRUST FOR THE WORKPLACE

The final stage of the zero-trust transformation is continuous improvement. Define and redefine the scope, equipment, and policies to keep pace with changes in devices, capabilities and organizational needs.

4.0

Summary

A zero-trust approach doesn't require a complete reinvention of your infrastructure. The most successful solutions should layer on top of and support a hybrid environment without entirely replacing existing investments.

Sharing dynamic context on identity, vulnerability and threat associated with users, devices and applications across all the various enforcement points is the best way to harmonize security policy, even though there will inevitably need to be different types of policy constructs and enforcement methods required to work with different parts of the environment.

Cisco Zero Trust

Cisco Zero Trust provides a comprehensive approach to securing all access across your applications and environment, from any user, device and location. It protects your workforce, workloads and workplace.

- **Duo** protects the workforce. With Duo's zero-trust workforce security, Cisco ensures only the right users and secure devices can access applications regardless of location.

- **Tetration** protects workloads. With Tetration's zero-trust workload security, Cisco secures all connections within your apps, across multi-cloud and the data center.

- **Software-Defined Access (SD-Access)** protects workplaces. Through SD-Access's zero-trust workplace security, Cisco secures all user and device connections across your network, including IoT.

This complete zero-trust security model allows you to mitigate, detect and respond to risks across your environment.

Learn more about **Cisco Zero Trust**.