zscaler™

White Paper

A brief history of zero trust: Major milestones in rethinking enterprise security

Why tell the story of zero trust?

Many in IT security believe zero trust is a game-changer, a fundamental rethink about enterprise security and protection of the networks and resources that house our best ideas, connect our brightest talent, and grant access to transformative productivity tools.

But to understand how truly revolutionary the zero trust model is in cybersecurity, it's necessary to understand the weaknesses of the legacy network-security approach and how the idea of zero trust architecture evolved into one that fundamentally overhauls decades-old thinking.

2D networks and castle-and-moat security

Hub-and-spoke and castle-and-moat are the two primary metaphors used to describe legacy network architecture and network security, respectively. Fittingly, the imagery used in both has been around for a while.

Hub-and-spoke network architecture refers to satellite networks arranged around a central hub. This model entails routing internal and external traffic back through a security stack at a primary data center before it proceeds to its destination. While this approach worked for a while, it's become more complicated and expensive given cloud adoption, distributed workforces, and the increasing importance of mobility in business.

Castle-and-moat security, on the other hand, refers to self-contained networks designed to admit friendly traffic while keeping enemies firmly outside their walls. Like a guard at the gate, in-house security appliances are meant to let the right people in while turning brigands away. The massive transition of applications to the cloud, coupled with the migration of workers outside of corporate perimeters, made this approach obsolete faster than cannonballs did for actual castles.

VPNs and Wi-Fi further complicated the problem. The old castle-and-moat architecture gave administrators no way to connect guests to a network without allowing them free rein while there. Ultimately, there was no good way to connect endpoints to networks without some form of segmentation to keep the latter secure.

We needed something better.

802.1X and the problems with NAC

In 2001, the IEEE Standards Association published its 802.1X protocol standard for network access control (NAC).

> "A means of authenticating and authorizing devices attached to a LAN port that has point-to-point connection characteristics, and of preventing access to that port in cases in which the authentication and authorization process fails."
>
> IEEE on 802.1X →

Soon after, wireless devices began to include an 802.1X supplicant, or client, that allowed networks to authenticate the endpoint before allowing a connection. This advance was intended to offer the ability to lock down wired and wireless networks, so that only managed devices and authorized users could connect. Think of the supplicant as providing ID to the bouncer at the network door deciding who's let in and who's left out in the cold.

Alas, the NAC model was no panacea – and the problems started with that N. internal networks were designed with implicit trust in mind, and trying to bolt on authentication/authorization after the fact was a huge effort. For NAC to be fully effective, all accessible ports needed to be locked down, but not all devices were 802.1X–capable. The rising adoption of internet–connected printers, badge readers, and other network–enabled devices was a glaring security hole. Now, imagine our bouncer was still manning only that one network door when multiple (or even dozens of) alternative entrances were available.

Toppling the walls of Jericho and rethinking the perimeter's role in security

By 2003, it was clear that personal device use would continue to proliferate, and organizations needed to start thinking about how to protect machines that weren't locked behind castle walls. Also, increasing use of encryption was reducing the effectiveness of perimeter firewalls, forcing a choice between scaling up to address capacity challenges imposed by decrypt–and–inspect, or allowing encrypted traffic to pass unchallenged.

That year, a multinational group of European technology leaders convened to address topics including user authentication, encryption, identity management, and policy enforcement. After formally establishing itself in 2004, the Jericho Forum introduced the notion of "de–perimeterization" to the world.

With a name recalling the Biblical story of the Israelites bringing down the walls of the ancient city of Jericho, the forum set about solving the problem of how to "enable secure, boundaryless information flows across firms."

In addition to the apt metaphor, the group left behind The Jericho Forum Commandments, the closest we'd come so far to truths from on high about governing perimeter–less networks. Unfortunately, the set of controls and mitigations prescribed was beyond the capability of most enterprises to deploy or administer at that point.

"Zero trust" first enters the IT lexicon

In 2010, Forrester analyst John Kindervag published a paper titled "No More Chewy

Centers: Introducing The Zero Trust Model Of Information Security" and, presto, we had a new buzzword representing a new way of thinking about network security. A key assertion of the paper was that the mere presence on a network was not sufficient for granting trust.

"This is where we started to hear things like, 'identity is the new perimeter,'" says Zscaler Field CTO and zero trust veteran Lisa Lorenzin. "We authenticated a user and used that identity to determine what they can do. Maybe, if we were lucky, we could gather some context like whether we had a managed or unmanaged device and make decisions about access based on that rudimentary understanding."

Progress. But this left enterprise security stuck on securing networks themselves. It wasn't yet ready to abandon them altogether. We were still falling short of a transformational approach, so the adoption of these principles again floundered. For one thing, it still relied on the same, network-focused toolset: 802.1X and RADIUS at Layer 2, identity-aware firewalls at Layer 3, etc.

The new way was just NAC with a catchy name.

Beyond(the perimeter)Corp

Meanwhile, hackers with ties to China's People's Liberation Army (PLA) were causing the tech industry's best and brightest to reconsider the issue of trust altogether. In 2010, Google disclosed a 2009 operation that had targeted it and several other high-profile tech companies including Akamai, Adobe, and Juniper Networks. The campaign was dubbed "Operation Aurora" by security researchers at McAfee.

By kicking the hornet's nest of elite IT engineering talent, Chinese hackers unwittingly accelerated work on zero trust architecture in the nation's top tech labs. Google developed BeyondCorp in response to Operation Aurora, which focused on **"shifting access controls from the network perimeter to individual users...[enabling] secure work from virtually any location without the need for a traditional VPN."**

But, "Google is a company run by engineers, for engineers, with an effectively infinite budget, and comparatively little legacy infrastructure compared to many enterprises," says Lorenzin. "And it still took them seven years and six white papers worth of design and implementation."

Even with Google's well-documented example, true zero trust architecture was still out of reach for most companies. Despite trying to "pave the path for other organizations to realize their own implementation of a Zero Trust network," the future Google imagined was still a ways off.

Meanwhile, for users, the popularity of the cloud and continued emphasis on mobility meant more data was available and being accessed from outside the network perimeter than within it. The need for a widespread approach to trust was greater than ever.

Gartner and the eventual arrival of zero trust network access

The tech research firm Gartner was responsible for the next significant advancements of zero trust as a broadly adaptable framework. While still around, the term "zero trust" wasn't top of mind in 2010 when the firm released its Continuous Adaptive Risk and Trust Assessment (CARTA).

The paper described the need to understand who's requesting access, and to grant that access based on a dynamic assessment of the environment, available context, and what a user's responsibilities warranted.

Lorenzin describes CARTA as “a great model that never got the traction it deserved.”

At Gartner, CARTA eventually morphed into “Zero Trust Network Access” (ZTNA) after the original framework failed to gain mindshare among tech practitioners (note the lingering focus on networks as the target of access!). But, fundamentally, CARTA remains important to the history of zero trust because the principles it laid out live on in the form of ZTNA.

Gartner’s next significant contribution to this discussion came with the recognition that the fields of networking and security were converging. In 2019, it expressed this marriage by introducing Secure Access Service Edge (SASE). It was a short-lived union, though, and by 2021 it was once again splitting the categories by introducing the Secure Service Edge (SSE) market category: SASE sans WAN.

Regardless of what you called it, Gartner had by this time established itself as a significant arbiter of what did or did not make for zero trust. Vendors were by now scrambling to fit themselves neatly into one of its new market categories.

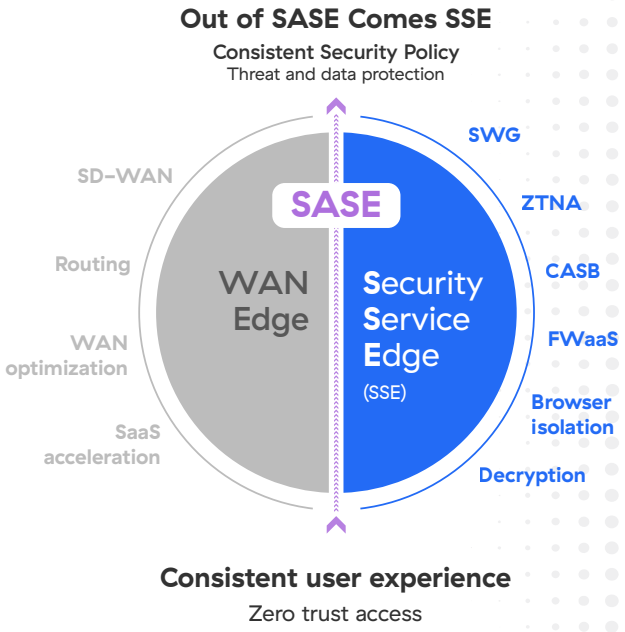
“The Man” enters the chat: NIST, OMB, and government endorsement of ZTA

In 2020, the National Institute for Standards and Technology (NIST) reframed the conversation with its NIST 800-207 standard for zero trust architecture. This new cybersecurity paradigm was focused on resource protection and the premise that trust should never be granted implicitly, but must continuously be evaluated.

With this paper, the shackles of the perimeter and the virtual private network were finally thrown off. The focus shifted from protecting the network to protecting the users, data, and applications interacting via the network. Zero trust now meant simply context-based, least-privileged access — applicable across a much wider variety of use cases and traffic flows.

The 800-207 standard stipulates key tenets and assumptions for zero trust. Three of the most critical points (from a much longer list) are:

1. No resource is inherently trusted.
2. All communication is secured regardless of network location. Terminating and inspecting the request; looking at all available context associated with the user and request.
3. All resource authentication and authorization are dynamic and strictly enforced before access is allowed.



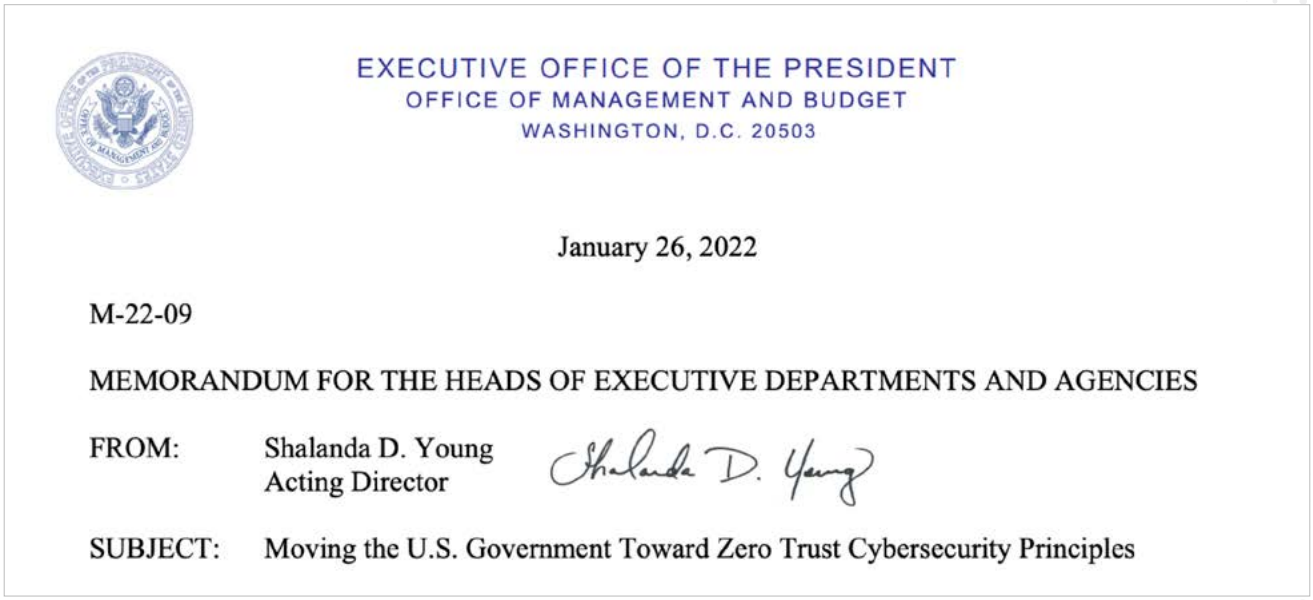
But the true point of no return for the promotion of zero trust principles came from the very top, at least in the United States. The U.S. Office of Management and Budget, the office responsible for implementing presidential policies, issued their M-22-09 directive in 2022, stating that all offices of the federal government must adopt zero trust architecture tenets by 2024 and outlining clear milestones and target dates along the way.

"So far, we've had guidance documents. We've had administrator models. But this is the first point where the rubber meets the road, with the federal zero trust strategy," according to Lorenzin.

The supply chain attack against the IT management platform Solar Winds — disclosed in 2021 and responsible for the compromise of at least nine federal agencies including State, Treasury, Homeland Security, Commerce, and Energy — was perhaps the most brazen and damaging state-sponsored attack since Operation Aurora. In response, the federal government has put its eggs in the zero trust basket, adopting that approach as its cybersecurity lodestar for the years ahead.

Implementing zero trust

Zscaler's approach to zero trust architecture aligns closely with NIST's ZTA framework and Gartner's definition for SSE. But it goes beyond any such standard, with its commitment to three fundamental advancements in zero trust thinking. Together, these advanced principles help push zero trust application to some logical conclusions.



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

January 26, 2022

M-22-09

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Shalanda D. Young
Acting Director

SUBJECT: Moving the U.S. Government Toward Zero Trust Cybersecurity Principles

All traffic is zero trust traffic

Zero trust began as a novel way to protect networks. Eventually, it expanded beyond on-premises networks but it was still focused primarily on private application traffic. For too long traffic was considered

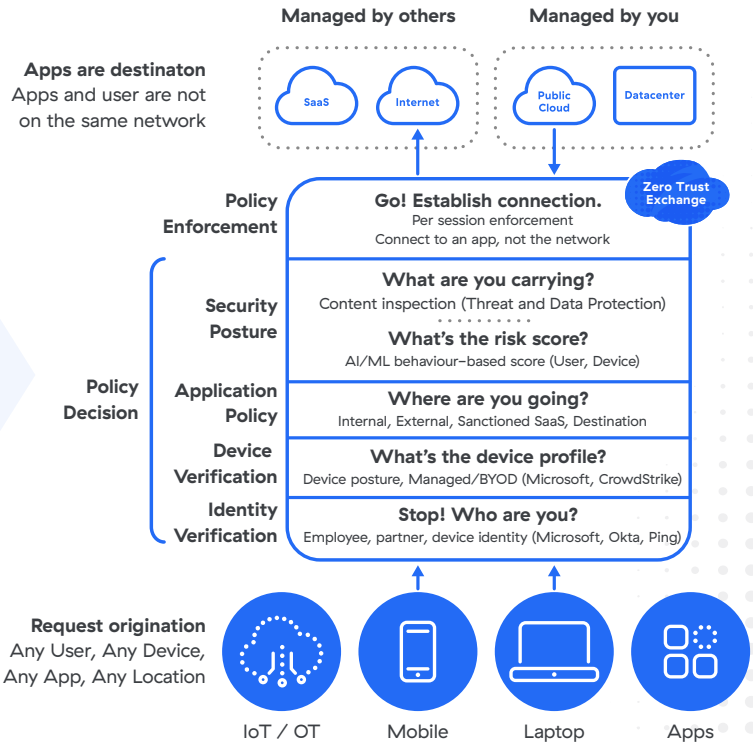
based on its relationship to a network, rather than doing away with the network altogether.

But we now know zero trust tenets can be applied to protect SaaS applications, traffic to and from public clouds, and even users as they access the public internet. And the originators of that traffic can be workloads, as well as users. Access can be made transport-agnostic, with traffic flowing via any router and coming over any network, wired or wireless, 4G or 5G, and so on.

It's past time to apply zero trust principles to all traffic, regardless of origin, regardless of destination. We've already done away with distinctions between trusted and untrusted, on-network or off. Now, it's time to stop thinking about what entity is connecting to what network, and instead use zero trust to connect all entities directly using business policies. The internet is the new corporate network and all traffic is fair game.

Zscaler Zero Trust Exchange Architecture

Per session policy decision and enforment across 150 DCs worldwide



1 **Identity and context always come before connectivity**

Identity verification lies at the heart of zero trust. But in the past we've confused identity with connectivity, and it's led us to broken models. IP addresses, MAC addresses, and port-and-protocol are not identity.

OT devices can connect to networks from factories. Users can log on from coffee shops. But that doesn't mean we know anything about them. So we have to start with identity and context. Only from there can we authorize connectivity.

When a user requests access to a resource, we must first consider who they are, other information about them such as role or department, the device they're using, and then security policies. What's the user trying to do? Where are they going? What in the environment might contribute to our decision to allow or deny the action?

Context goes beyond identity and is evaluated continuously. Other factors that can be cross-checked for anomalies include geolocation, IP address, device posture, and time of day. And a zero trust solution should be able to decrypt traffic, to inspect for threats and data exfiltration risks inline and at scale.

2 In the case of the Zero Trust Exchange, we also correlate threat intelligence — from across our global cloud, as well as from third-party technology partners like security and identity verification vendors — to determine risk and make policy and access decisions.

Applications – and even app environments – should remain invisible to unauthorized users

Now that we've solved the problem of knowing who you are before granting you access, we can tackle the next challenge: how do we connect you to your authorized resources, while reducing risk and minimizing the potential for compromise? Once the context surrounding a user, device, policy, and environment has been gathered and analyzed, we can take the next steps in that direction.

By eliminating the inbound listener for remote connections, we eliminate the external attack surface. Otherwise, it's simply too easy for attackers to locate vulnerable VPN gateways or exposed applications to compromise targets. VPNs sitting around awaiting inbound connections are sitting ducks, and threat actors take notice. This is a vendor-agnostic problem — it can only be solved by changing the architectural model.

The Zscaler Zero Trust Exchange does this by forming outbound-only connections both from the user and from the application environment out to our security cloud using encrypted micro-tunnels to broker connections between requests and their destinations.

3 This online "third place" provides a buffer between verified users and any resource they are authorized to access. Once a user is connected to the requested asset, granular policies ensure that there's no option to venture beyond it. Lateral movement becomes essentially impossible.

The final chapter?

The principles discussed above allow us to truly and finally move past a legacy understanding of network perimeters guarded by firewalls, and remote endpoints connected via virtual private networks. They don't just replicate existing security controls in a cloud-hosted virtual instance, or rely on some artificial understanding of what's on the network versus what's not.

A comprehensive architecture designed to deliver zero trust security — for users, workloads, applications, OT and IoT devices, and beyond — reduces risk, improves protection, simplifies the user experience, and represents a fundamental improvement in the way we think about enterprise security.

+1 408.533.0288 Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134 zscaler.com