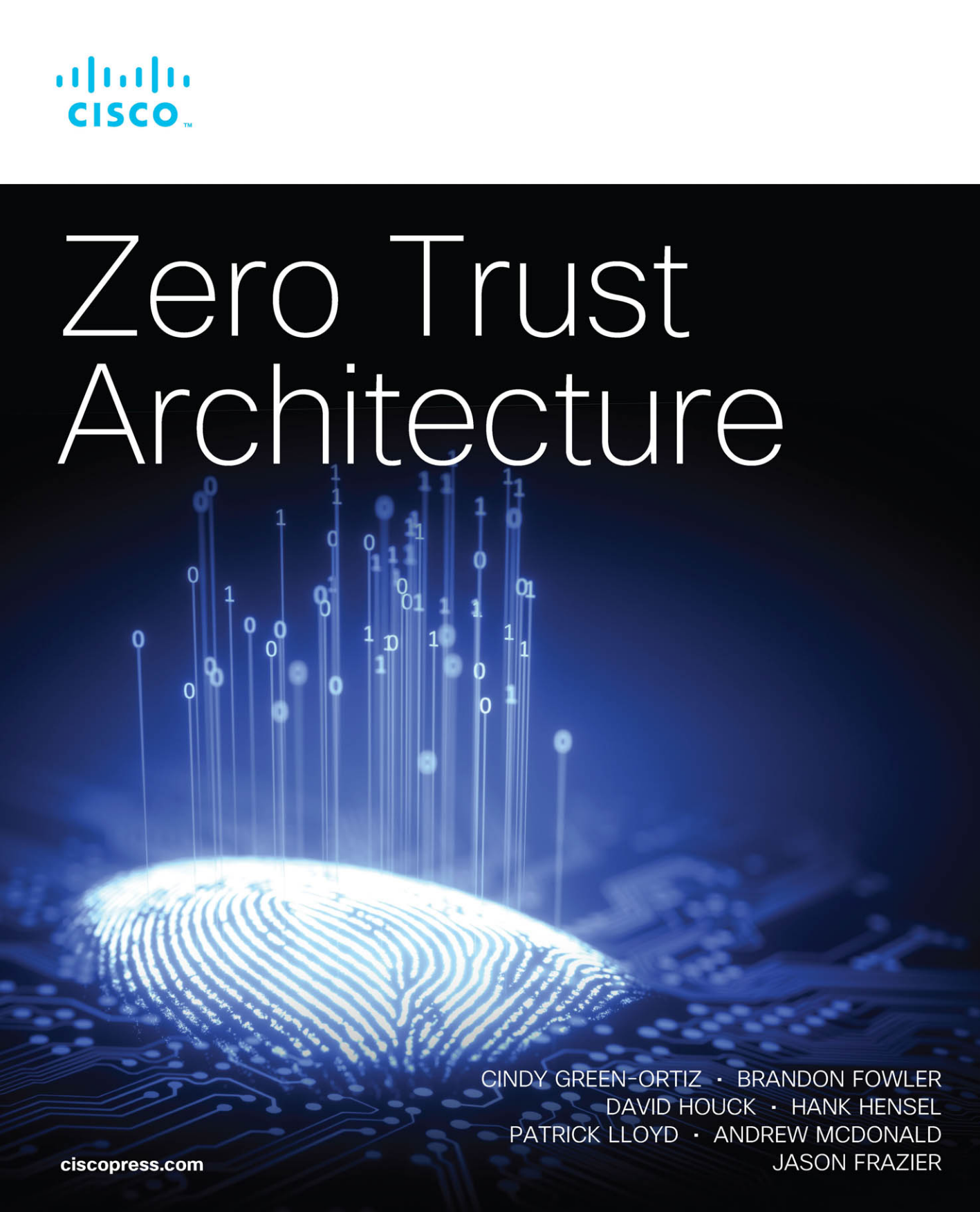
Zero Trust Architecture
CINDY GREEN-ORTIZ · BRANDON FOWLER
DAVID HOUCK · HANK HENSEL
PATRICK LLOYD · ANDREW MCDONALD
JASON FRAZIER
ciscopress.com

Zero Trust Architecture

Cindy Green-Ortiz, CISSP, CISM, CRISC, CSSLP, PMP, CSM
Brandon Fowler, CCNP Security
David Houck
Hank Hensel, CCIE No. 3577, CISSP
Patrick Lloyd, CCIE Enterprise No. 39750, CISSP
Andrew McDonald
Jason Frazier, CCSI

Cisco Press

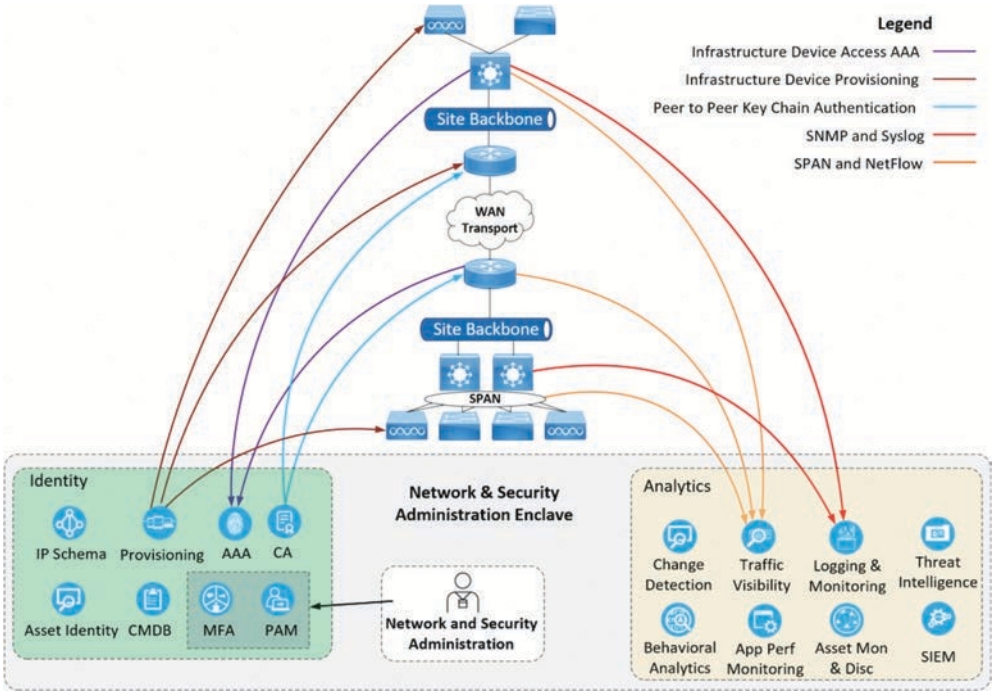

Figure 3-5 *Network Telemetry*

WAN

Unlike the branch or campus network, the WAN is a more hands-off area when it comes to Zero Trust. That said, the WAN continues to have many of the same concepts of Zero Trust as the branch and campus networks, as illustrated in Figure 3-6. This stems from the WAN being in one of two models, either network access devices that are owned by an organization and terminating circuits that are typically leased to them from a service provider, or the model of a fully owned and managed WAN by a service provider. In the case where equipment is owned by an organization, the same concepts apply as did the network:

- Identity of the network access device being synchronized with the policy application/enforcement server.
- TACACS+ being used for authentication of actions being taken on the device.
- Command-by-command authorization when actions are taken.
- Integration into a SIEM for tracking all changes made.
- As-a-service offerings from service providers or partners including web application firewalls or distributed denial of service protection may be available and more suitable than on-premises implementations.

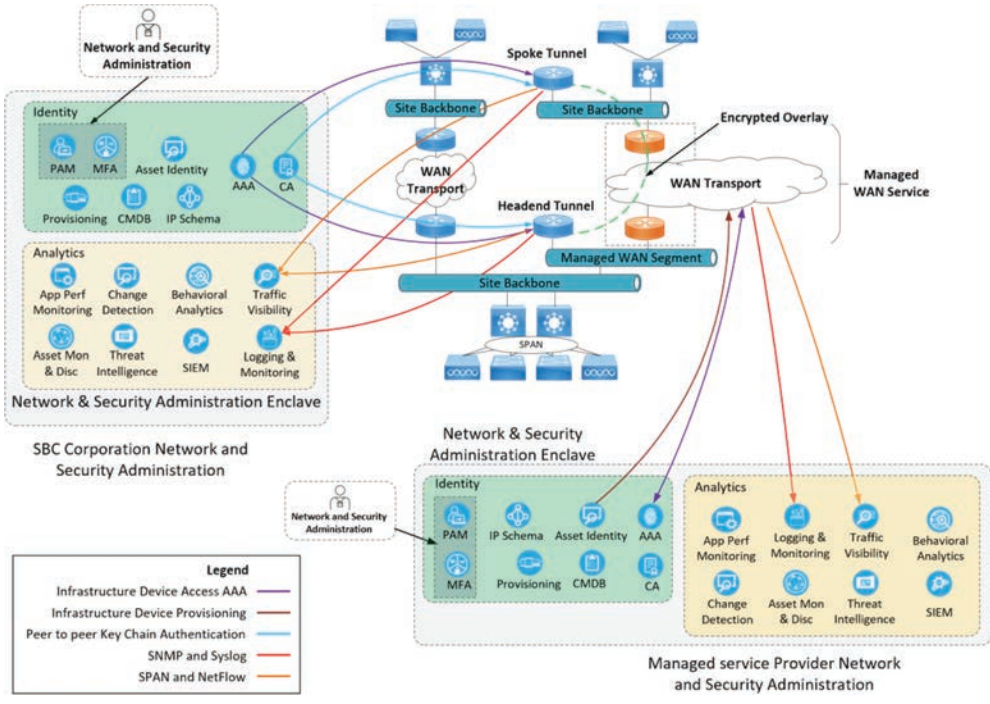

Figure 3-6 *WAN Control Points*

Where a Zero Trust application for the WAN skews from branch and campus networks, the recommendation is to utilize an overlay to protect traffic as it traverses the WAN. This is especially true when the WAN is fully managed and operated by another entity. The greatest source of concern for an organization when it comes to its WAN is the potential for a man-in-the-middle attack. Man-in-the-middle attacks on the WAN are facilitated by the flow of packets through the WAN provider's infrastructure and the owning organization having little visibility of the traversal of that data. In the same way that NetFlow and network taps were recommended for branch and campus networks, WAN providers may use the same mechanisms to understand packet flow and troubleshoot traversal across the WAN for customer data flows. Given the likelihood of secured protocol traffic being decrypted when built into applications, having a mechanism to universally encrypt all traffic in transit is highly recommended. Utilizing an implementation of SD-WAN, such as the Cisco SD-WAN series of implementations, provides the added benefit of carrying segmentation data in the packet as well, creating a full fabric where policy can be applied.

A fabric overlay, implemented over the top of the WAN, such as Dynamic Multipoint Virtual Private Network (DMVPN), Group Encrypted Transport Virtual Private Network (GETVPN), or IPsec VPN configured in a full mesh allows for securing of traffic flow. Depending on the protocol, exposure of the sender's IP, the tunnel source's public IP, or an overlay-based IP can provide significant flexibility in what informa-

tion is exposed while in a service provider's cloud. As a bonus to encrypting the links and preventing man-in-the-middle–type attacks, when utilizing segmentation tagging technologies, like Cisco TrustSec, the TrustSec tag can be written directly into the DMVPN, GETVPN, or IPsec tunnel packet, and allow for that information to traverse the WAN. Without these technologies the information from the tag may very well be stripped. This allows for policy application that occurred at one site to traverse the WAN, allowing for additional identifiable information for endpoints to become ubiquitous throughout the network.

Data Center

The data center has always been the nerve center of most organizations. It is where the "crown jewels" are typically stored, and where most of the major servers and applications that run the business and process the data critical to business success exist. Therefore, it only makes sense that a major focus needs to be put on determining what exists in the data center and how it communicates within and external to the data center, and validating that endpoints in the data center belong there. Commonly, devices will be hosted in data centers as opposed to other areas of the architecture, due to the "free" ability to power, cool, and maintain the devices away from scrutiny within smaller branches and campuses. However, there have also been numerous examples of unauthorized servers existing in data centers, hosting P2P file-sharing activity, hosting websites for nation-states, crypto mining, and other exploitative activities where data centers are not the focus of security.

While unauthorized usage should be a major concern for data centers and their operators, the potential for a seemingly innocuous and otherwise authorized endpoint to be used within the boundaries of the data center that then infects the data center through no purposeful fault of the user is an even greater concern. As mentioned in the introduction to this section, securing a data center has always been a relatively straightforward task, when it focused on threats being on the outside of the data center and needing to find a hole to get in. However, when servers exist in the data center that may not be authorized, or even when authorized may host malicious software, apps, or data, it puts an organization at as much, if not more, risk than that from external attackers.

As a prime example of this, the IRS found 1150 servers within its data center of 1811 that were unauthorized, and the threat they presented to the network was immeasurable. One reason why threats from the outside were easier to secure against is the traffic traversal nature of servers communicating to the outside world needing to communicate through a firewall. The firewall then logs at least an IP, hopefully static in nature, that can be used to track activity of the server. When it comes to internal threats posed by virtual servers, the lack of a definitive hardware enforcement point prevents an easy understanding of communications.

While unauthorized usage should be a major concern for data centers and their operators, the potential for seemingly innocuous and otherwise authorized endpoints to be

used within the boundaries of the data center also poses a significant risk. As mentioned in the introduction to this section, securing a data center has always been a relatively straightforward task, when it focused on threats being on the outside of the data center and needing to find a hole to get in. However, the common nature of small, portal, fully functional operating systems that can easily be connected without significant visible scrutiny also poses a significant threat to the network. Where servers exist in the data center that may not be authorized, there is the potential to host malicious software, apps, or data, putting the organization at risk as much as, if not more than, that from external attackers.

The challenge with data centers is the mixed nature of the data center and how the continued evolution of the data center impacts what can be monitored and how. The desire to condense data centers, minimizing cooling and power costs, has caused many organizations to use various virtualization models to collapse multiple physical servers into a single physical server. A "hypervisor" or virtual management plane utilizes a single or dual physical network interface card to send traffic for all the respective virtual servers on the physical chassis. While each server may have its own IP address, the lack of ability to apply policy to a physical network port prevents granular policy from being applied. While there may be some ability for the hypervisor itself to be a centralized control point, integrations between the virtual and physical architectures still have many gaps. The inability to track communications of virtual servers within the same hypervisor, or even within the data center, limits the enforcement abilities to limit these workload-based threats. This limitation may be overcome through agents installed on the servers or with hypervisors that act like switches, with the ability to apply policy to the unique virtual machine "sessions" correlated to the virtual network interface card.

Therefore, when it comes to data center topologies related to Zero Trust, many of the same methods apply that have been referred to in this book thus far, just with the additional challenge that the virtualized nature of the data center may impact the ability to fulfill a lot of the need for discovery and enforcement. A sample of a typical data center architecture may be found in Figure 3-7. To prevent against unauthorized servers having access through the data center network, the first requirement of a machine that connects to the data center network is a clear identity that can be presented to the network and used to track activity of endpoints contained therein. The challenge in a data center related specifically to identity is that most servers will either not have a user logged in to them or will have a service account that is used to maintain the login, while individual users are given specific rights to make changes or configure the server respective to their application's needs. This means that a server may either have no specific user identity presented to the network, and therefore needs to consider other aspects of its contextual identity to continue to build the identity it uses to validate its business purpose on the network, or that the same physical or virtual server may have many identities utilizing it, and therefore present multiple identities over a short period of time. Regardless, this contextual identity needs to be aggregated and aligned with the device for analysis purposes.

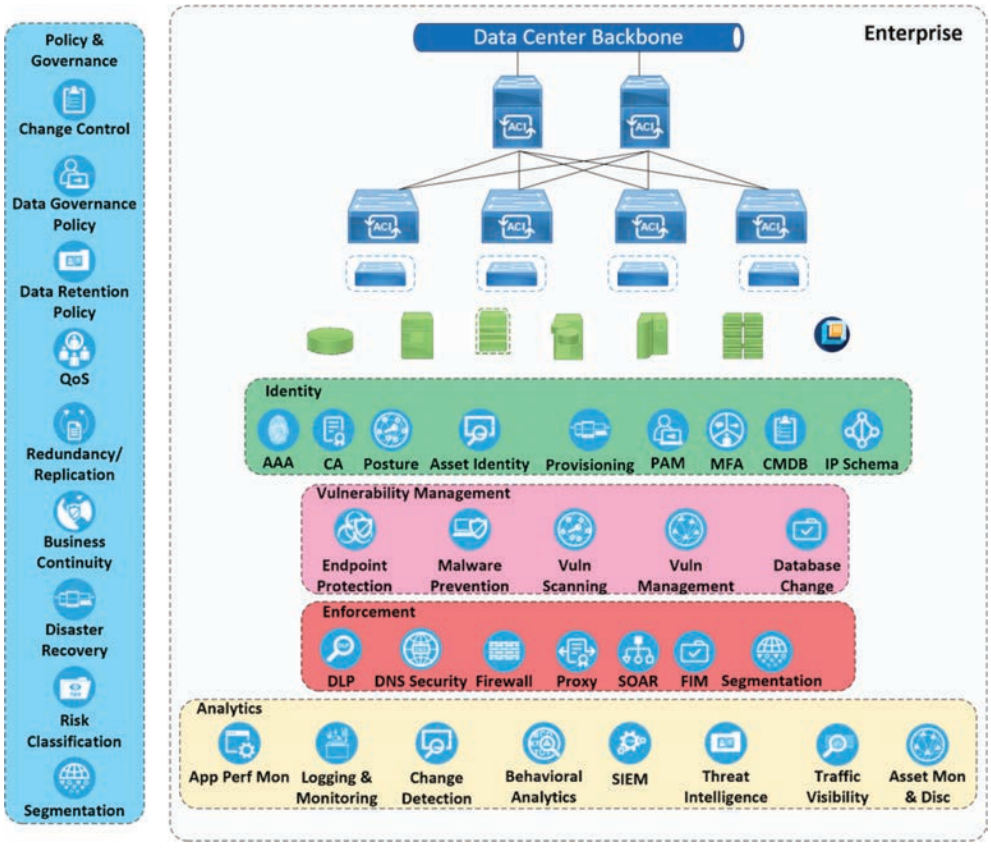

Figure 3-7 *Data Center Architecture*

The second challenge that may happen in the data center regarding endpoints relates to vulnerability management. Once a server can be accurately identified with a contextual identity, it then needs to be validated in terms of its posture and hardening status from internal attacks. While some vendors have made hardening servers a menial task with group policies that can be replicated based on identity, the validation of anti-virus, anti-spyware, or anti-X can be a challenging conversation in many environments due to the perception that running these utilities could slow the processing time of valid workloads, or worse yet, not be supported due to the age of the server. In many environments, monolithic servers and applications live on in infamy due to their massive cost to replace or rewrite the applications residing upon them. This means out-of-support operating systems that no longer have anti-X applications written for them, and for those that can be worked around to force anti-X to work on them, a lack of ability to centrally manage those applications due to their age and own support models.

In these situations, a business case may be able to be made for a longer-term success, which includes rewriting the application, or segmenting the server in such a way that

duplicates with its same functionality exist and can process records should that server be compromised while not communicating with one another in such a way that would compromise the shared workload if one were infected in some way, shape, or form. A shared back-end database with read replica-type models, or copies of data synchronized in a snapshot type model, may be of use to ensure that potential compromise in this environment would not inhibit business as usual. In addition, good practices would include asset management policies that define appropriate configurations for out-of-compliance systems such as legacy servers and operating systems, which will help to mitigate risk by having rigorously reviewed hardened configurations and standards.

Once this final challenge is overcome, multiple enforcement mechanisms exist, dependent upon the type of server in question. For virtual servers, both in private data centers as well as public clouds, the use of an on-server enforcement agent or a policy-based gateway is required to apply policy due to the inability of the hypervisor to recognize dynamic policies applied to the switchport itself. This is due to the RADIUS enforcement mechanism, based on a common session ID between the network access device and RADIUS server, with a need to apply that policy to the specific virtual machine in question. This policy gateway could be in the form of another virtual machine that acts as the network access device itself, such as a virtual switch that is onboarded into the RADIUS server, or could be in the form of an aggregation point that all traffic is sent through, which applies policy to all devices. Where a centralized policy enforcement agent was not feasible to the data center design or throughput requirements, agents can be deployed to individual servers, which act as micro enforcement agents and are pushed policies for what traffic the server can send and receive.

For physical servers, the enforcement mechanism requires particular care to be demonstrated in ensuring that data center switches being used support Zero Trust-level enforcement. For many environments, vendor production of switches has focused on specific use cases, like a trading floor with micro-second or pico-second latency or a large data center with massive throughput needs with very few "bells and whistles." With any of these use cases, the addition of security mechanisms to be included in the feature set of the switch may not have been a consideration, or special configuration circumstances may exist. There can be situations when utilizing RADIUS on these servers that there may not be support for a change of authorization. Another case is when using a tagging mechanism such as TrustSec, the tag may need to be statically assigned to the port, port profile, VLAN, or subnet.

Regardless of how enforcement and policy application are done within the data center, overlay and analytics principles hold as they would with any other topology. There is a need for understanding of traffic flow, along with regulatory requirements that apply to the data that exists on the endpoint and how enforcement is applied to the endpoint, regardless of the enforcement mechanism. Logging and analytics of access attempts, in addition to the flow of traffic to and from the server and associated identities of that traffic, need to be aggregated, reviewed, and validated for any potential intrusion or threats to the network. Please find a sample Cisco ACI Fabric Policy Model used to provide Zero Trust Controls in a data center in Figure 3-8.