CMMC AND SPLIT TUNNELS

Authors: Ace Swerling, Ryan Bonner, Richard Wakeman

Solutions for the Cybersecurity Maturity Model Certification Practice SC.L2-3.13.7

Contents

Introduction

Cybersecurity Maturity Model Certification (CMMC) is the DoD's effort to improve the safeguarding capabilities of defense contractors throughout the defense industrial base (DIB). CMMC extends the control set required by Defense Federal Acquisition Regulation Supplement 252.204-7012 (DFARS 7012), which currently drives contracting organizations to implement National Institute of Standards and Technology (NIST) Special Publication (SP) 800-171 to protect Controlled Unclassified Information (CUI) generated under Department of Defense (DoD) contracts.

The NIST SP 800-171 standard required by CMMC 2.0 includes an obligation to avoid split tunnel Virtual Private Networks (VPNs). This prohibition has caused confusion for organizations pursuing such VPNs based on recommendations from their technology vendors. This paper intends to address this confusion by explaining the nuances behind the controls and the technology from the perspective of the-security community.

Control Discussion

The control prohibiting split tunnel VPN is SC.L2-3.13.7, which reads "*Prevent remote devices from simultaneously establishing non-remote connections with organizational systems and communicating via some other connection to resources in external networks* (i.e., split tunneling)."

Organizations can be confused about exactly how and when the split tunneling prohibition applies based on its wording. To avoid this, it is important to understand the control prohibits split tunnels only when connecting to services that are *outside* the Controlled Unclassified Information (CUI) boundary. It is acceptable to configure a split tunnel to a cloud service that has been configured to demonstrate compliance with NIST SP 800-171 and/or CMMC 2.0 maturity level 2 or higher.

For example, a workstation residing on the internet may directly communicate (split tunnel) to both an enterprise datacenter and an internet-based (cloud) service so long *as both are within the CUI boundary* and have been configured to NIST SP 800-171. Conversely, this direct communication would *not* be allowed if the cloud service, for example, was not in the CUI boundary nor configured to NIST SP 800-171.

Put another way, the control allows split tunneling to any asset, including cloud services, if they are defined as in-scope within an organization's SSP. Split tunneling is disallowed when connecting directly to services out of scope of the SSP.

This point will be discussed in detail later.

Split Tunneling Background

Split tunneling is a networking technique that sends some traffic over a VPN and other traffic through another route. This is considered a split tunnel VPN because the traffic is split between the two locations using two tunnels. This scenario often arises when somebody's work requires securely accessing resources in both a internal enterprise services and internet-based cloud services.

Optimal networking efficiency calls for routing traffic directly to the resources, be it the datacenter or the cloud service. The traffic is protected by encryption and the user has a much better experience. An alternative is to route all traffic from a workstation through a centralized network device and then allow traffic to route as needed from there, also known as a hairpin through a VPN chokepoint. They both get the job done, but the hairpin introduces performance issues as the network traffic must travel further to reach destinations outside the centralized environment.

While it might be tempting to use split tunneling to mitigate performance issues when connecting to cloud services, if not configured properly, this approach may leave web traffic vulnerable to cybersecurity threats that are intended to be addressed by NIST SP 800-171.

The tunnels are secure if properly configured and are not inherently concerning. The concerns leading to Control SC.L2-3.13.7 arise since many organizations use their external network devices (like firewalls and VPN concentrators) as monitoring points for network traffic, looking through network flow information for malicious activity.

> These monitoring points are also generically called *Policy Enforcement Points* (PEPs) and are called *Managed Access Control Points* in NIST 800-171. CISA defines these in the TIC 3.0 Reference Architecture as "security devices, tools, services, or applications that enforce the security capabilities. Enforcement may occur at any point between endpoints. Enforcement actions include permit, deny, modify, redirect, delay, and other forms of data manipulation. The actions are initiated based on a variety of attributes, as defined in security policies."

When using a split tunnel, the network traffic moving between an endpoint and a cloud service can bypass the network monitoring and policy enforcement if it does not traverse through the external network endpoints, i.e. the PEPs. This results in a lack of visibility into security threats, which motivated NIST's prohibition of split tunnel VPNs. To mitigate this, the NIST control requires all traffic from a

remote workstation to traverse a VPN tunnel terminated at a location managed by the organization so it may be monitored for security issues.

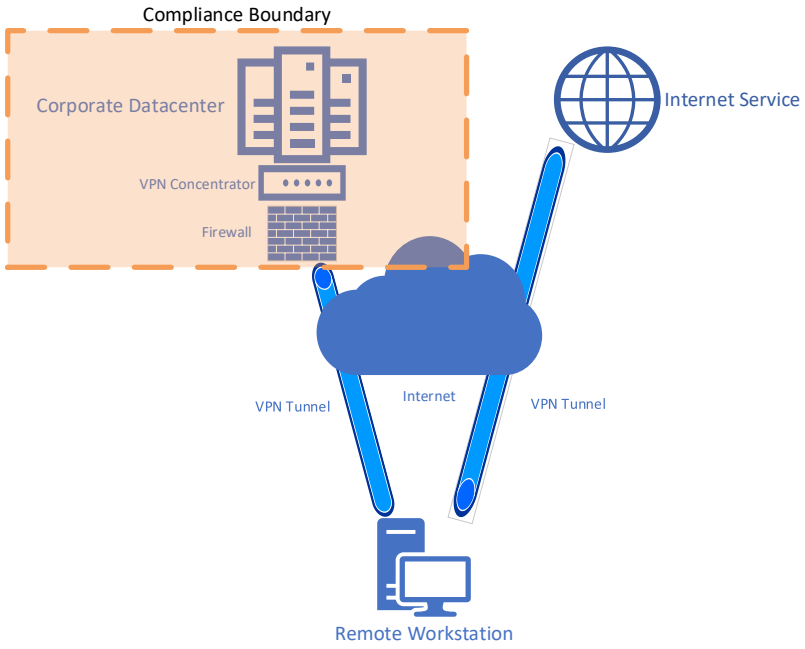

Figure 1 – Split Tunnel VPN Prohibited by CMMC Controls

However, this type of VPN tunnel routing can significantly impact network performance and cost, especially in latency-sensitive applications like audio/videoconferencing (Teams, Zoom, GoToMeeting, etc.) as traffic to these services must route through the enterprise network over the VPN tunnel.

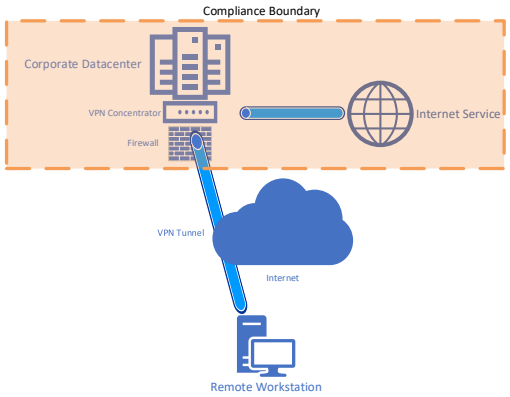

Figure 2 – No Split Tunnel as Required by CMMC Controls

The additional network overhead introduced by this configuration is often barely noticeable when browsing the internet but can introduce a poor user experience to applications that are sensitive to network timing and jitter like voice and videoconferencing. This dissatisfaction can lead to users to seek alternative means of collaboration, often outside of corporate governance, such as unmonitored phone calls.

This performance problem led leading cloud service providers (CSPs) like Zoom and Microsoft to recommend traffic directly connect with their cloud services to avoid these latency issues.

This places the optimal technical approach at odds with security requirements, creating a dilemma between the performance demanded by users and security demanded by the regulations.

Solutions for Cloud Services

There are several options to resolve the dilemma between efficient consumption of online services and safe monitoring of their use. Moving security services to the cloud allows organizations to reroute Internet traffic through a ubiquitous security layer in the cloud while continuing to rely on VPN protection for traffic flowing in and out of the on-premises datacenter.

1. Move To the Cloud

Split tunneling may be avoided altogether if there are no resources in a corporate-owned network, which would require no VPN. For example, using Microsoft 365 for collaboration (including Teams for conferencing) and Azure for computing places everything within a single technical environment and CUI boundary. Placing computing resources within Amazon AWS and using their Chime service or using Google's GCP, Google Docs, and Google Meet could achieve the same goal, *assuming all services reside within the same compliance boundary.* Any connection to a service outside the boundary would be considered split tunneling, prohibiting it from holding CUI.

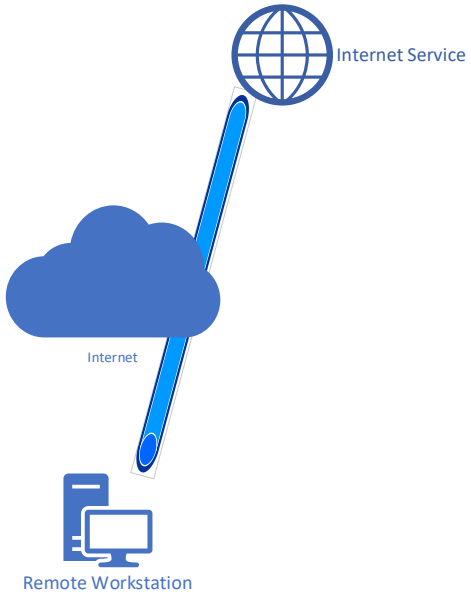

Figure 3 – Everything in The Cloud = No Split Tunnel

2. Hybrid-environment with Dynamic Routing

Another option is to include all cloud services within a single CUI boundary, applying NIST SP 800-171 VPN controls to both the datacenter and any other internet-based service. This will adequately protect the data as required by the regulations because it will always be covered by the security controls.

In these cases, split tunneling is allowed because the prohibition only applies to connection to *external* services, i.e. those that are outside the CUI boundary. This makes sense as the intent for the SC.L2-

3.13.7 control is to protect CUI by ensuring all network traffic is monitored. Since all services within a CUI boundary are adequately monitored and the CUI never passes outside the organization's visibility, the data stays protected even when running a split tunnel VPN.

The split tunnel is achieved via what may be called "*dynamic routing*" or *"hybrid VPN"* by configuring a conditional access rule on the organization's VPN. Traffic from a trusted endpoint to a cloud service is allow-listed, permitting traffic to bypass a VPN device in the enterprise datacenter and communicate directly with the *cloud* services. This functionality is known to be provided by VPNs from Fortinet, Cisco, Palo Alto, Azure, and Watchguard, although it is not limited to these vendors' products.

Walking through an example, say somebody uses Microsoft 365 on a workstation and uses Teams for videoconferencing. Also say this person frequently works from home and accesses data within an enterprise datacenter. This user would need to access the datacenter via a VPN to facilitate this secure access and protect the data as it transits the internet.

It is recommended to configure the VPN to recognize when traffic is intended for the organization's CMMC-compliant Microsoft 365 environment. In this case, the VPN would see Microsoft 365 on its allowlist and instruct the VPN software residing on the workstation that it may connect directly to Teams without VPN to operate as efficiently as possible.

Conversely, that user may receive an invitation from an outside vendor to join a call on Zoom. The Zoom service is outside the CUI boundary and does not fall under the organization's CMMC boundary. This makes Zoom an external service as defined by 800-171 and CMMC, meaning the prohibition on split tunneling applies. The VPN service would not see Zoom on the allowlist so all traffic would need to traverse the enterprise datacenter via the VPN before it can reach Zoom. The same situation would apply for any other internet service, including Gmail, Salesforce, or Facebook.

Note: all companies and services are mentioned as examples. There is no limitation inherent to any of these services. Instead, the limiting factor is whether they are configured consistent with the security frameworks, allowing them to reside within an organization's compliance boundary.

More details on this approach may be found in Matt Titcombe's blog post at CMMC, Split Tunneling, and COVID | Peak InfoSec.

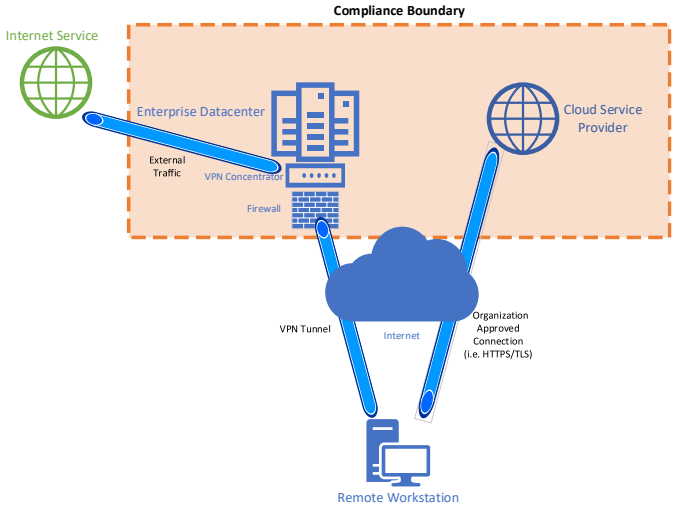

Figure 4 – Dynamic Routing Is Allowed Within A CUI Boundary

3. Zero Trust Architecture

Zero Trust Architecture (ZTA) has been gaining additional attention within the US government per an Executive Order from U.S. President Biden. At its core, ZTA evolves organizations away from the historical network-based boundaries in favor of boundaries around individual users and resources. This is achieved by tying access to a user's identity, not the network or the data's placement within it. ZTA denies access by default and only grants access if conditions like the correct user account, access rules, device health, and risk measures are met.

ZTA brings much greater flexibility and power to protecting data while requiring changes to IT architectures and security frameworks. ZTA does not force an organization to choose between it and traditional network controls. Instead, ZTA is often introduced incrementally to a traditional network. ZTA should be considered an addition and improvement to network protection, not a replacement.

ZTA is relevant to split tunnel scenarios because this evolution presents the opportunity to secure CUI without having to traverse a VPN concentrator residing at a network perimeter.

US government standards consider ZTA an emerging space as shown in Zero Trust Architecture (nist.gov), ZTA Reference Architecture (defense.gov), and CISA TIC 3.0 Reference Architecture v1.1. It is expected these references will be used to guide implementations in response to the Executive Order, but at the time of this document's release, specific technical details of the government's implementation expectations have not been published. At the time of publication, this means NIST SP 800-171 and CMMC are not completely aligned with ZTA although it is expected they will be revised to better reflect the government's migration toward ZTA. In the meantime, this lack of clarity from the US Government makes an *exclusively* ZTA-focused approach a risky path to compliance. Even if ZTA can result in effective security, the divergence with the regulations will motivate many auditors to raise issues.

Organizations may consider implementing ZTA along with traditional perimeter-based protections, including firewalls and VPN concentrators, while the government works to clarify its approach and evaluation criteria under CMMC. There is substantial security benefit to organizations from ZTA and it's unnecessary to delay implementation until the government determines how to audit this. Recall that the end goal for is to secure data and businesses. Audits and certifications are means to this end, not the goal, which means it is more important to focus on securing the data in a compliant way rather than pursuing solutions to meet a compliance target.

With that said, what's special about ZTA and why is it relevant to split tunnels?

Zero Trust Changes the Game

Organizations used to establish trust by working from secure facilities, controlled areas, and on organizationally controlled networks. Security was enforced at the edge of a corporate network, defining a perimeter at junctions between an organization's network and the internet. This was considered a boundary between an "internal" and "external" network. The internal network would be trusted because it was entirely within an organization's control and the external network would be untrusted because it was not.

Organizations would place a firewall at these junctions to control and monitor traffic traversing this boundary. The firewall is hardened software that determines what network traffic is allowed to flow between the internal and external network. The theory is that this restriction will keep out malicious activity, allowing less rigorous security on software and services on the internal network.

The reality is that computing has changed, threats have become more sophisticated, and firewall-defined boundaries don't support scenarios like joint ventures very well. Scenarios such as working from home are hastening this change since employees are no longer working in static environments. Firewalls still have a role to play but it's unwise to rely on them as heavily moving forward.

Instead of relying on network security services like firewalls and VPNs that restrict network flow, ZTA uses a person's identity and the rules pertaining to data access to determine what's allowed. These credentials are checked before any regulated data is allowed to flow. If a user and his/her device are allowed access to a bit of data, then the traffic will be allowed to flow. If not, then nothing will be allowed. This deny-by-default security posture ensures there is no standing access to applications nor resources, especially for elevated privileges.

To make this happen, the network protection that was previously applied only at a network perimeter is shifted to all devices under an organization's control so that no network is considered trustworthy. There is no "internal" or "external" network. Instead, the system only considers "do you have access or not?"

This changes the nature of the CUI boundary, which has traditionally focused on protecting data on the "internal" network.

Zero Trust Uses New Technology

The ZTA concept encrypts data at rest and in transit while allowing network activity monitoring. In other words, with ZTA there is still a centralized Policy Decision Point (PDP) but enforces protections through distributed Policy Enforcement Points (PEPs) and devices' traffic tunnel across all networks (even if the network is corporately managed) through encryption effectively protecting all traffic as a VPN would. This avoids both security risks and performance issues pertaining to split tunnels.

This is achieved by implementing a combination of protective capabilities on computing devices, especially those that may reside outside an organization's traditional network perimeter. These include, but are not limited to:

1. Dynamic Routing: ZTA will still use conditional routing with a hybrid VPN similar to Option 2. This capability allow-lists services on the internet allowing direct connections from a workstation. This effectively routes all traffic to an organization's perimeter-based VPN endpoint except for services where a direct network connection is desired and allowed.
2. Endpoint Detection and Response (EDR): This software is installed on workstations and servers (endpoints) and allows distributed application of network access policy from a centralized console. It also looks for threats on these endpoints and reports to the centralized console so action may be taken. This moves the security monitoring function from the network perimeter to endpoints, ensuring sufficient coverage as required by the regulations.
3. Antimalware: This software is installed on endpoints and protects against viruses, worms, and other kinds of malware. Reporting will be forwarded to a centralized console, similar to EDR, so security operations may respond to malware wherever an endpoint may be.
4. Data Loss Prevention (DLP): This software encrypts data at rest based on classifications and detects if it leaves a protected environment. This capability applies an additional level of protection, especially to regulated data, that is otherwise difficult to achieve. DLP defines a logical boundary based on the scope of data protection. This logical boundary can be applied regardless of physical or organizational boundaries, which helps support the latest highly abstracted multi-cloud and multi-organization architectures.
5. Mobile Device Management (MDM): This service manages endpoint configuration, including for workstations and mobile devices (smartphones, tablets, etc.) This software allows devices to roam outside the organization's network perimeter but still be securely configured. MDM can be used to enforce items like device encryption, password policy, and minimum patch levels. MDM also includes device-level authentication so only recognized devices may access data. In this way, devices are handled similarly to people and service accounts. Prohibiting connections from unvalidated devices helps protect against malicious activity.
6. Device Authentication: This capability ensures a device is sufficiently secure before allowing it to access sensitive information. It can check whether the device is sufficiently patched, has been encrypted, requires secure authentication, and has updated antimalware signatures. This is another form of validation, helping to ensure devices are safe before accessing any data.
7. DNS Filtering: This service works like an antimalware or spam filter for DNS queries. It maintains a blacklist of addresses that are known to be malicious, which helps block the internet-based command and control networks that are often used by ransomware and other malware.

8. Risk and Location-Based Filtering: This service monitors network connections and may prohibit connectivity to or from untrusted locations. This may include blocking network connections to embargoed locations (e.g. Section 126.1 Nationals) or with undue risk, such as from impossible locations.

Zero Trust Shifts the Scope of CMMC

As a certification, CMMC is validates an assessment boundary. What happens when your assessment boundary doesn't look like a traditional network and security perimeter?

Assessment boundaries used to be the logical networks themselves, with user accounts and devices roaming freely inside the permitter from one network to another once traffic rules granted access. Zero trust tools create perimeters around individual resources, and enforces rules based on a composite (attribute based) algorithm for policy enforcement. Because of this shift, the perimeter shrinks, from an IP range or domain, down to a unique user, device, geolocation, and secure system state. When this happens, the notion of a point-to-point VPN disappears because all network traffic is secured as if it were outside of a traditional perimeter and on a VPN. This eliminates the need for VPN tunnels altogether.

Zero Trust Consolidates CMMC Practices

As the boundaries for implementation (and CMMC assessment) become smaller: zero-trust architectures create defense-in-depth scenarios where certain CMMC practices, written for a broad spectrum of system designs, begin to satisfy practice objectives related to other CMMC practices.

The zero-trust concept of micro-segmentation creates an assessment scenario where the stacking of multiple practices directly satisfies the practice objectives of other CMMC practices. When combined with encryption practices for data in transit and session authenticity (3.13.8, 3.13.11, 3.13.15), all outbound connections from devices within the organization's zero-trust boundaries are also controlled (satisfying SC.L2-3.13.7).

In a ZTA environment, full tunnel VPNs are replaced by host-based protections including Endpoint Detection and Response (EDR) and TLS-encrypted network connections. EDR technology implements a centralized framework for network policy that is applied to remote devices. This creates a centrally managed and monitored network protection capability as required by the CMMC controls while distributing the protection across all computing devices, not only the perimeter-based firewalls. Meanwhile, the TLS-encrypted network connections protect the data in transit.

As a reminder, the objective for practice SC.L2-3.13.7 requires that "remote devices are prevented from simultaneously establishing non-remote connections with the system and communicating via some other connection to resources in external networks (i.e., split tunneling)." In a zero-trust architecture, other CMMC practices satisfy this practice objective. In a design where networks are not considered trustworthy, host-based firewalls provide boundary protection on the device itself. Because CMMC 2.0 Level 3 requires organizations to adopt a deny-by-default posture to all inbound traffic, non-remote (local) connection, attempts from external networks will be dropped by the host-based firewall rules. CMMC 1.0 Level 3 also required the implementation of DNS filtering (SC.3.192), requiring all outgoing

requests to utilize the DNS filtering service. This has been deprecated with the introduction of CMMC 2.0, making it no longer required, but is still a good security practice.

Additional Controls Covered by Zero Trust

As zero-trust architectures become commonplace, other examples of CMMC practice consolidation are becoming more evident:

- Device-based certificates and hardware-based trusted platform modules (TPM) asserted as credentials in device authentication (SC.L2-3.5.1, SC.L2-3.5.2) provide a higher degree of trust when compared to username/password combinations that can be used from untrusted systems.
- Using the secure state (or "health") of a device as a trust factor when deciding to grant access to resources (SC.L2-3.1.2) can provide logical access restrictions (SC.L2-3.4.5) for privileged users seeking to make system administration changes.
- Identifying the geolocation of users and devices can be used to establish new external boundaries (SC.L2-3.13.1, SC.L2-3.13.5) in lieu of traditional facilities and networks.
- Dynamically adjusting users' access privileged based on their current trust algorithm represents the real-time enforcement of least privilege (SC.L2-3.1.5) and least function (SC.L2-3.4.6).

Concluding Summary

ZTA has been accepted by the US Department of Defense as a valid technical approach that provides sufficient security assuming, of course, that it is properly implemented and maintained. ZTA is not yet reflected in the NIST 800-171 controls, although an update is presumed to be coming.

Organizations can benefit from ZTA regardless of the government's plans for CMMC certification because of its utility and power to protect business and data from substantial threats, including data theft and ransomware.

The point is to protect CUI and maintain a CUI boundary. Whether your approach includes dynamic routing, ZTA, or anything else is up to you.

The first two approaches discussed in this paper – avoiding split tunnels or implementing dynamic routing – are reasonable in the near term and are the most straightforward paths to NIST SP 800-171 compliance. They may not be practical, performant, or usable though. ZTA is the IT industry's most flexible and direct approach to respond to the latest threats. Organizations will be well served by including ZTA in their security plans, especially those that are already undertaking remediation efforts. The recommendation is to pursue Option 1 (split tunnel avoidance) or Option 2 (dynamic routing) in the near term while simultaneously implementing Option 3 (ZTA) with a view to the future.