A dangerous game of hot potato

crowe.com/cybersecurity-watch/ntlm-relay-attacks

NTLM relay attacks are common. Organizations should know what they are and how to protect against them.

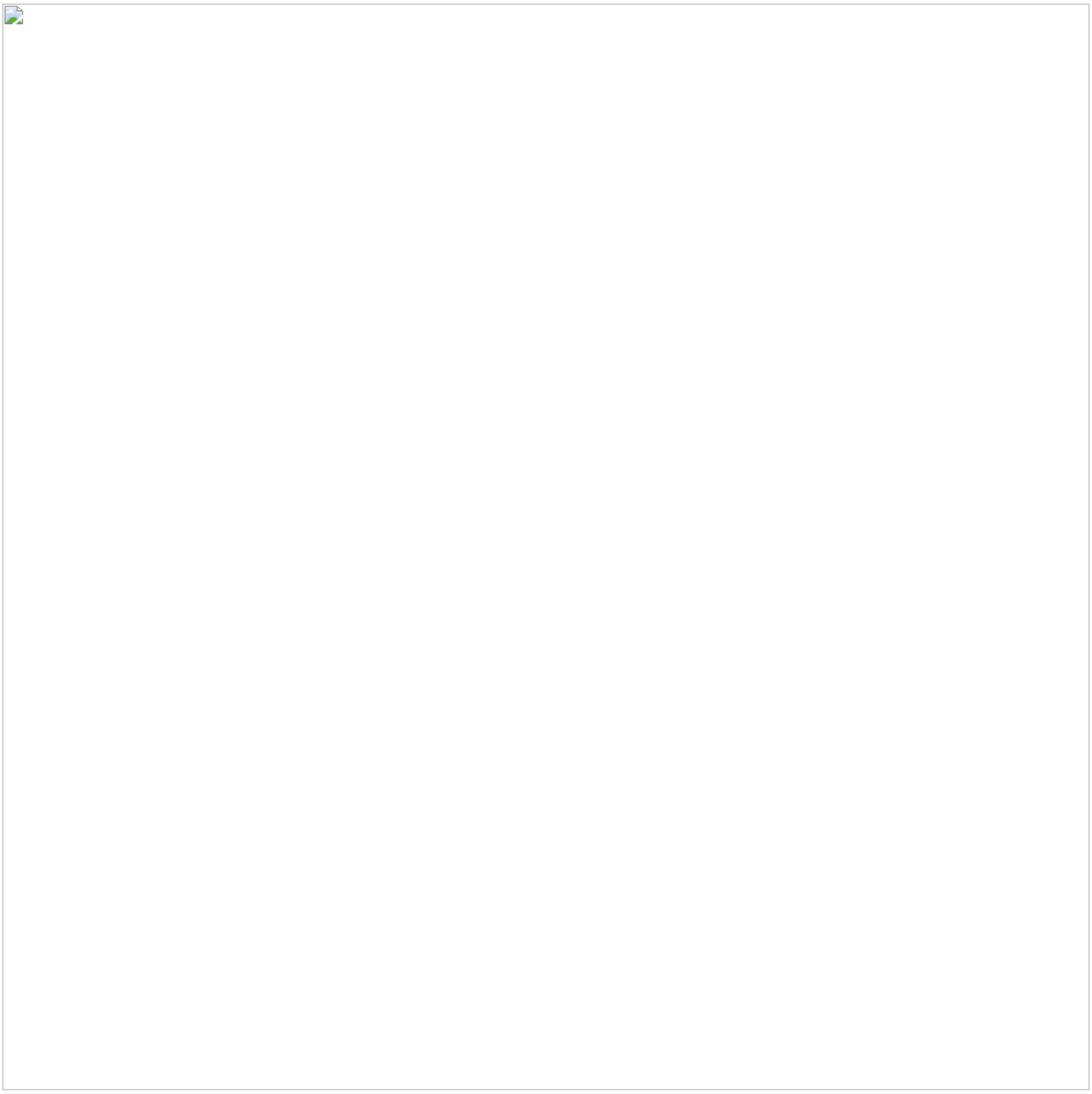
In a Microsoft Windows™ environment, authentication is often synonymous with Windows New Technology LAN Manager (NTLM). Despite being replaced as the primary authentication protocol by Kerberos in Windows 2000, NTLM remains pervasive even in new environments. NTLM continues to be used because of the need for backward compatibility and because it is still enabled by default in Windows 10 and Windows Server 2019.

NTLM relay attacks use old authentication protocols that make 1980s-type assumptions about trust, and they grant access in the process. Organizations should learn how these attacks work and then take steps to protect against the many forms these attacks can take.

What is a relay attack?

When compromising an environment, gaining network access is only half of an attacker's challenge. Sensitive resources are often secured from outsiders through authentication, leaving only affiliated devices with the ability to gain access. Attackers dream of simply having to ask an affiliated device to share access – relay attacks provide exactly that.

How relay attacks work



The purpose of relay attacks is to redirect authentication from one source to another. An attacker can trick a system (Device A) into authenticating to an attacker-owned machine. Thus, the attacker can relay that authentication (like a hot potato) to a target system (Device B) and harness the valid credentials to use them on an unintended system.

Vulnerable by design

Many of NTLM's vulnerabilities are intrinsic and can't be patched without fundamental changes to its design. NTLM uses a challenge-response protocol to authenticate users when the client and server devices perform operations on shared secrets and exchange three messages. When a user attempts to access a resource, NTLM authentication typically follows the steps specified below:

- The client conceals its password using a one-way hash function.
- The client sends its plaintext username to the server.
- The server replies with a random number as a challenge.
- The client sends the challenge encrypted using its hashed password.
- The server repeats the client's encryption using its known correct password.
- The server compares its answer to the client's and grants access if they match

Though NTLM has undergone upgrades since its LAN Manager origins in the late 1980s through NTLMv1 and NTLMv2, these upgrades primarily focused on increased challenge length and tougher hashing algorithms. The fundamental challenge-response mechanism has remained the same, meaning the inherent vulnerabilities are present unless additional mechanisms are used. Microsoft has issued limited updates to patch the numerous vulnerabilities associated with NTLM, but this patching is only akin to applying simple bandages over severe wounds.

Attack basics

Relay attacks have a long tradition of use, and one of the earliest implementations was SMBRelay in 2001 by Sir Dystic of the hacking group cDc. This implementation relied on the resource sharing protocol server message block (SMB) and followed a similar architecture to more recent relay attack methods:

- The attacker sets up a malicious relay purporting to provide a form of service.
- The victim client is tricked into attempting to access the service
- The attacker service relay demands the client authenticates to receive access.
- The client follows the typical authentication process and sends its proof of authenticity to the attacker.
- The attacker relays the received messages at each step to a legitimate server to gain access itself. More recent implementations of the attack allow for multiple relays of the same credentials. Alternatively, the attacker can stash hashes away for later cracking.

The issue with NTLM and other protocols that makes relaying possible is the lack of mutual authentication. While NTLM's challenge-response mechanism is simple and enables a server to authenticate a client desiring access, it leaves the client inherently trusting the server. Like many older and simpler protocols such as NetBIOS and LLMNR, an assumption is made that the other devices in the network segment are legitimate.

Recent NTLM relay attack implementations

When attackers find themselves in a Windows environment containing NTLM clients waiting with anticipation to send out their credentials, the objective typically is to identify a means to force the client to authenticate to the attacker – that is, to figure out a nice way to "ask" for credentials. Rather than relying on typical poisoning attacks, attackers can find a specific vulnerability to exploit to coerce authentication on demand. These means have been the subject of many more recent implementations of NTLM relay attacks including two examples: PetitPotam and DFSCoerce.

- **PetitPotam**. Released by Topotam in 2021, PetitPotam is a novel means of forcing a client to authenticate. It takes advantage of Microsoft Encryption File System Remote Protocol (MS-EFSRPC) to convince a victim to authenticate over Microsoft Local Security Authority Remote Procedure Call (MS-LSARPC) on port 445. The attack can be performed either with or without authentication, depending on the API command used (RpcRemoteFindFirstPrinterChangeNotification versus EfsRpcOpenFileRaw). Once attackers receive the victim's authentication, they can redirect the authentication to gain access to a more critical server, such as Microsoft Active Directory™ Certificate Services (AD CS). Doing so enables attackers to assign themselves a security certificate that allows them to gain access to a domain controller (DC) and thus the entire Windows environment.
- **DFSCoerce**. DFSCoerce is newer exploitation in the same family as PetitPotam; it was released in 2022 by Wh04m1001. Instead of MS-EFSRPC, it uses Microsoft Distributed File System Namespace Management (MS-DFSNM) to force a DC to authenticate against an NTLM relay. Authentication is similarly redirected at an AD CS server over HTTP to get a Kerberos ticket-granting ticket that allows device impersonation. This redirection effectively allows an attacker to move from being a domain user to a domain admin without even requiring permission to use the DFS service.

Detection

Both PetitPotam and DFSCoerce are easily detectable with proper logging due to their unique events and network traffic. PetitPotam attacks can be identified based on anonymous logins and suspicious connections to SMB shares as well as by the use of elevated tokens from authenticating with a machine account. DFSCoerce can be identified through the attempted additions and removals of DFS namespaces. Looking at network traffic, both attacks can be found through their suspicious relaying behavior as well as the unusual DFS and SMB connections.

Remediation

Because attackers can always discover new attack primitives to trigger authentication, the best way to eliminate NTLM relay attacks is to disable NTLM on DCs in favor of Kerberos. The use of mutual authentication and the requirement that a client be domain-joined with access to a DC prior to access make the protocol far less susceptible to relay attacks. However, Kerberos is not a silver bullet and still can be relayed under specific conditions.

Kerberos relay attacks

Kerberos authentication requires the use of an encrypted service principal name (SPN) that identifies the target of authentication. An authentication request for an attacker’s service can’t simply be relayed to authenticate to another device’s service. Therefore, attackers must force a victim client to generate an authentication request for an SPN different than the one it is connecting to. Accomplishing this step is complex, and it demands specific conditions because it relies on additional protocols and applications such as Lightweight Directory Access Protocol (LDAP), web browsers, internet protocol security authentication, Microsoft Remote Procedure Call, or domain name system. In many cases, a single flag being set can stop an attack, limiting the scope of Kerberos relay attacks. Most attacks follow this general pattern:

- The attacker responds in a way that specifies the SPN of the desired authentication target.
- The victim client uses the specified SPN to make an authentication request sent to the attacker.
- The attacker relays the authentication to the desired authentication target to gain access.

Additional mitigations

Luckily, NTLM and Kerberos share mitigations that add additional layers of security. Such additional layers include required SSL/TLS and extended protection for authentication, which require that authentication occurs over the client’s initial transport layer security connection, as well as SMB and LDAP signing, which confirm the authenticity and origin of packets.

Defeating NTLM relay attacks

Relay attacks can come in many forms, but the bottom line is that they can effectively subvert authentication on internal networks. Understanding how the attacks work is critical to determining how best to stop them. Mitigative measures, especially Kerberos, can introduce additional complexity to an environment and could cause backward compatibility issues. However, not protecting against these attacks likely will cause even larger headaches.

If implementing Kerberos is too far of a jump, SMB and LDAP signing can block many common relay methods. Most importantly, taking the time to understand common attacks and establishing layers of security can help protect an environment against relay threats.

Microsoft, Active Directory, and Windows are trademarks of the Microsoft group of companies.