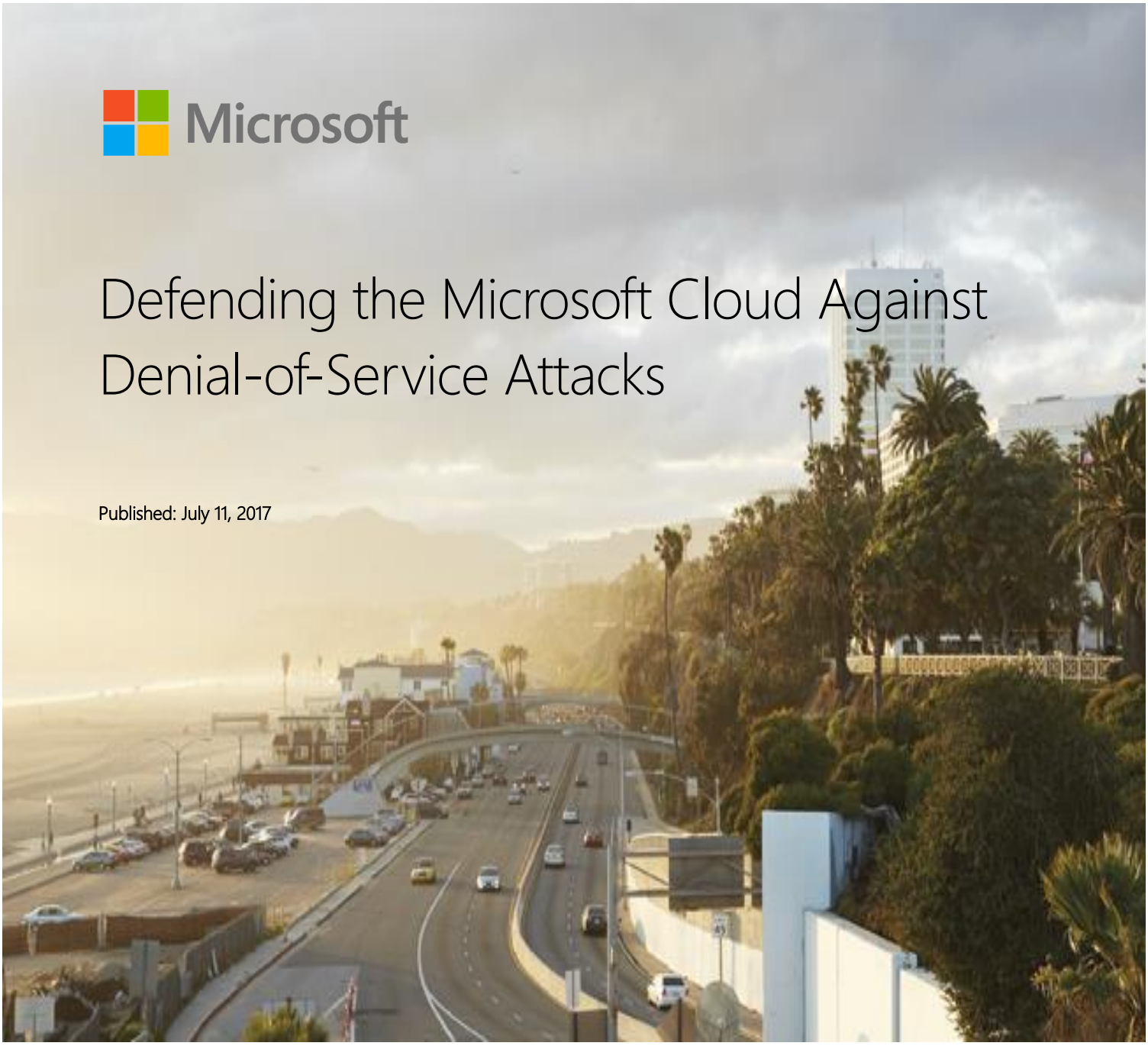

Defending the Microsoft Cloud Against Denial-of-Service Attacks

Published: July 11, 2017

This document describes general types of network-based cyberattacks and how Microsoft defends its cloud services and networks against them

Introduction

Microsoft delivers a trustworthy infrastructure for more than 200 cloud services, including Microsoft Azure, Microsoft Bing, Microsoft Office 365, Microsoft Dynamics 365, Microsoft OneDrive, Skype, and Xbox Live that are hosted in our global cloud infrastructure of more than 100 datacenters.

As a global organization with a significant Internet presence and many prominent Internet properties that provide cloud services, Microsoft is a large, common target for hackers and other malicious individuals. The network--the communication layer between clients and the Microsoft Cloud--is one of the biggest targets of malicious attacks. In fact, for many years, Microsoft has been continuously and persistently under some form of network-based cyberattack. At almost all times, at least one of Microsoft's Internet properties is experiencing some form of attack. Without reliable and persistent mitigation systems that can defend against these attacks, Microsoft's cloud services would be offline and unavailable to customers.

Microsoft uses defense-in-depth security principles to protect its cloud services and networks. This document describes general types of network-based cyberattacks and how Microsoft defends its cloud services and networks against them.

Definition and Symptoms of Denial-of-Service Attacks

One way to attack network services is to create many requests against a service's hosts to overwhelm the network and servers to deny service to legitimate users. This is referred to as a denial-of-service (DoS) attack. When the attack is performed by multiple actors, endpoints, and/or vectors, it is referred to as a distributed denial-of-service (DDoS) attack. Although the means, motives, and targets vary, DoS and DDoS attacks generally consist of the efforts of a person or persons to prevent an Internet site or service from functioning correctly or at all, either temporarily or indefinitely.

The United States Computer Emergency Readiness Team (US-CERT) defines symptoms of DoS attacks to include:

- Unusually slow network performance (when opening files or accessing Internet sites)
- Unavailability of a Web site
- Inability to access a Web site
- Dramatic increase in received spam
- Disconnection of a wireless or wired Internet connection
- Long-term loss of access to the Web or any Internet services

Overview of Attacks

Network-based cyberattacks manifest in five primary ways:

- Bytes/sec (bps)
- Packets/sec (pps)
- Transactions/sec (tps)
- Connections/sec (cps)
- Maximum concurrent connections (mcc)

Bytes/sec Attacks

Fundamentally, a bytes/sec (bps) attack is about sending more data than the network can handle. It focuses on saturation based on the size of the data as opposed to the rate of transmission. The goal is to send the network so much traffic that it starts discarding packets. Fundamentally the attacker is trying to saturate a fixed link of a determined size between two devices. The malicious traffic consumes so much of the available bandwidth that legitimate requests can no longer be sent over the saturated link. The payload that is transmitted is often random and irrelevant.

By way of example, one common method of attack is to use Network Time Protocol (NTP) reflection. An NTP reflection attack is one in which the attacker sends a small amount of data to an NTP server with a spoofed IP address (the victim's IP). This causes the responding NTP servers to send large amounts of traffic in the form of responses to these requests to the victim's IP address(es), thereby swamping the victim's network. Another example is an attack that broadcasts an Internet Control Message Protocol (ICMP) ping containing the victim's IP address. This form of attack amplifies the bps attack, thereby increasing the amount of traffic sent to the victim.

Packets/sec Attacks

This attack exploits the simple fact that not all packets are created equal. It focuses on using an excessive number of packets to cause saturation, versus using a large amount of data to cause saturation. There is a fixed cost for processing packets, regardless of the size of the packet. The bandwidth capabilities of a given network assume well-formed traffic (e.g., 1500-byte packets). Even very fast networks and network devices can slow down when their packet-per-second limits are reached. Smaller size packets (e.g., 64 bytes) can cause bandwidth maximum capacity to drop significantly.

Instead of causing a lot of traffic to overwhelm a network or device, an attacker instead sends a lot of small packets, thereby extending the time it takes to process the packets and thereby slowing down the network. UDP flood attacks are a common way of conducting this type of attack.

Transactions/sec Attacks

This type of attack is tailored to its intended target and often happens after previous Bytes/sec and Packets/sec attacks fail to make a service or services go offline. The attacker(s) will analyze the service

that is being run and then attempt to perform transactions against that service's components to see how fast they respond. The attacker is trying to figure out which transactions have a longer response time because that indicates the service is performing more work and consuming more resources to respond to the request. If the attacker has a good understanding of the target's architecture, they can further tailor the attack, and with just a few packets they can disrupt service.

For example, say an attacker knows that a service will consume 10% CPU time when any of the following activities are performed:

- Three concurrent search operations
- Seven concurrent successful logins
- One re-indexing operation
- Four concurrent login failures

In this example, an attacker would simply need to send 40 concurrent bad login requests per second to consume 100% CPU time on the target. Well before the Bytes/sec and Packets/sec thresholds are reached, the target has been taken offline.

Attackers commonly perform significant pre-attack probing when issuing transaction-based attacks, often using bots (a software application that runs automated tasks over the Internet). Because of Microsoft's Digital Crimes Unit works both in cyberspace and with the justice system, Microsoft has successfully disabled significant numbers of bots, while at the same time building a vast database of known bots and infected IP addresses. Microsoft shares this information with customers with Azure Active Directory Premium so that they can perform their own forensic correlations and analyses.

Connections/sec Attacks

This attack involves attempting an extremely high number of connections to overwhelm the devices receiving the requests. By filling the devices' connection tables, new connections are refused, and thus legitimate users of the site are unable to use the service. A common method of attack is a SYN flood attack where an attacker sends a succession of SYN requests to fill up the connection table. SYN auth is a common and inexpensive protection mechanism used to defend against SYN flood attacks.

Maximum Concurrent Connections Attacks

Like Connections/sec, this attack is targeted against devices that maintain state and connection tables. But whereas Connections/sec attacks focus on the rate of new connections, Maximum Concurrent Connections attacks focus on the total number of connections, often generating these connections slowly as to avoid detection, and then keeping them open as long as possible. Slowloris is a hacker tool that is commonly used to conduct these attacks.

Core Principles of Defense

The three core principles when defending against network-based DoS attacks are Absorption, Detection, and Mitigation.

Absorption happens before detection, and detection happens before mitigation. Absorption is the best defense against a DoS attacks. If the attack can't be detected, it can't be mitigated. But if even the smallest DoS attack can't be absorbed, then services aren't going to survive long enough for the attack to be detected.

Of course, it is generally not economically feasible for most organizations to purchase the excess capacity necessary to absorb DoS attacks, as this requires a considerable investment in technology and technical skills. This highlights one of the security benefits of using Microsoft cloud services; the sheer scale of our services enables us to provide strong network protection to our cloud customers in a cost-effective manner. But even at our scale, though, there must still be a balance between absorption, detection, and mitigation. To find that balance, we study an attack's growth rate to estimate how much we need to absorb.

Detection is a cat-and-mouse game. You must constantly look for the new ways people are attacking you or trying to defeat your systems. Detect -> Mitigate -> Detect -> Mitigate, etc., is a perpetual, persistent state that will continue indefinitely.

Defending Against DoS Attacks

To successfully defend against a DoS attack, early detection is essential. By detecting an attack before the system is overwhelmed, defenders can execute a response plan.

The following formula will help approximate the time to impact of a DoS attack:

$$\text{Maximum Capacity} / (\text{Maximum Capacity} \times \text{Growth Rate}) = \text{Time to Impact}$$

If the time-to-detection occurs after time-to-impact, then it is likely the DoS attack will be successful. If the time-to-detection occurs before time-to-impact, then the services being attacked should remain online and accessible, if mitigation strategies are used. Thus, there are only two things that can be done to defend against DoS attacks:

1. Increase capacity to raise the ceiling of maximum capacity (which in turn provides more time to detect an attack); or
2. Decrease the time to detect.

Increasing capacity has a direct fiscal impact. Microsoft recommends that customers develop at least basic absorption capacity, to ensure that they can survive some level of DoS attack. The actual absorption capacity will vary from customer to customer, as each customer has their own thresholds for exposure, risk, and financial outlay. Ultimately, for economic reasons, investments of research and time in ways to decrease time-to-detection are usually the most cost-effective defense.

Microsoft's DoS Defense Strategy

Microsoft's strategy for defending against network-based DoS attacks is somewhat unique due to our scale and global footprint. This scale allows Microsoft to utilize strategies and techniques that few organizations (providers or customer organizations) can match. The cornerstone of our DoS strategy is leveraging our global presence. Microsoft engages with Internet providers, peering providers (public and private), and private corporations all over the world, giving us a significant Internet presence (which as of this writing, doubles around every 18 months). Having such a large presence enables Microsoft to absorb attacks across a very large surface area.

Given our unique nature, Microsoft uses detection and mitigation processes that differ from those used by large enterprises. Our strategy is based on a separation of detection and mitigation, as well as global, distributed mitigation through our many edges. Many enterprises use third-party solutions which detect and mitigate attacks at the edge. As our edge capacity grew, it became clear that the significance of any attack against individual or particular edges was very low. Because of our unique configuration, we have separated the detection and mitigation components. We have deployed multi-tiered detection that enables us to detect attacks closer to their saturation points while maintaining global mitigation at the edge. This strategy ensures we can handle multiple simultaneous attacks.

One of the most effective and low-cost defenses employed by Microsoft against DoS attacks is to reduce our attack surface. Doing so enables us to drop unwanted traffic at the edge, as opposed to analyzing, processing and scrubbing the data inline.

At the interface with the public network, Microsoft uses special-purpose security devices for firewall, network address translation, and IP filtering functions. We also use global equal-cost multi-path (ECMP) routing. Global ECMP routing is a network framework that ensures there are multiple global paths to reach a service. Thanks to these multiple paths, an attack against the service should be limited to the region from which the attack originates – other regions should be unaffected by this attack, as end users would use other paths to reach the service in those regions. We have also developed our own internal DoS correlation and detection system that uses flow data, performance metrics and other information. This is a hyperscale cloud service running within Microsoft Azure which analyzes data collected from various points on Microsoft networks and services. A cross-workload DoS incident response team identifies the roles and responsibilities across teams, the criteria for escalations, and the protocols for engaging various teams and for incident handling. These solutions provide network-based protection against DoS attacks.

Finally, cloud-based workloads are configured with optimized thresholds based on their protocol and bandwidth usage needs to uniquely protect that workload.

Defending Microsoft Cloud Services against DoS Attacks

Microsoft datacenters are protected by defense-in-depth security that includes perimeter fencing, video cameras, security personnel, and secure entrances that use biometrics, smartcard, and multi-factor authentication. The defense-in-depth security continues through every area of the facility and to each physical server unit. The Microsoft Cloud Infrastructure and Operations Group delivers the core infrastructure and foundational technologies for our cloud services. Our datacenters comply with industry standards for physical security and reliability and are managed, monitored, and administered by Microsoft operations personnel.

To further protect our cloud services, Microsoft provides a DDoS defense system that is part of the Microsoft Azure continuous monitoring and penetration-testing processes. The Azure DDoS defense system is designed not only to withstand attacks from the outside, but also from other Azure tenants. Azure uses standard detection and mitigation techniques such as SYN cookies, rate limiting, and connection limits to protect against DDoS attacks.

Microsoft's cloud services are subject to the threat of attack from multiple sources. To mitigate and protect against the various DoS threats, a highly-scalable and dynamic Azure-based threat detection and mitigation system has been developed and implemented with the primary objective of protecting the underlying infrastructure from DoS attacks and helping to prevent service interruptions for cloud services customers. The Azure DoS mitigation system protects inbound, outbound, and region-to-region traffic.

Most DoS attacks launched against targets at the Network (L3) and Transport (L4) layers of the Open Systems Interconnection (OSI) model. Attacks directed at the L3 and L4 layers are designed to flood a network interface or service with attack traffic to overwhelm resources and deny the ability to respond to legitimate traffic. Specifically, L3 and L4 attacks attempt to either saturate the capacity of network links, devices, or services or overwhelm the CPUs of servers or virtual machines supporting an application.

To guard against L3 and L4 attacks Microsoft has designed, developed, and deployed a solution aimed specifically at safeguarding the infrastructure and customer targets that come under attack by protecting these layers. Protecting the infrastructure ensures that attack traffic intended for one customer does not result in collateral damage or diminished network quality of service for other customers. The solution uses traffic sampling data from datacenter routers. This data is analyzed by a network monitoring service to detect attacks. When an attack is detected, automated defense mechanisms kick in.

Application-Level Defenses

Microsoft engineering teams follow the rigorous standards set by Microsoft Operational Security Assurance to help protect customer data. Microsoft's cloud services are intentionally built to support a very high load as well as to protect and mitigate against application-level DoS attacks. We have

implemented a scaled-out architecture where services are distributed across multiple global datacenters with regional isolation and throttling features in some of the workloads.

Each customer's country or region, which the customer's administrator identifies during the initial setup of the services, determines the primary storage location for that customer's data. Customer data is replicated between redundant datacenters according to a primary/backup strategy. A primary datacenter hosts the application software along with all the primary customer data running on the software. A backup datacenter provides automatic failover. If the primary datacenter ceases to function for any reason, requests will be redirected to the copy of the software and customer data in the backup datacenter. At any given time, customer data may be processed in either the primary or the backup datacenter. Distributing data across multiple datacenters reduces the affected surface area in case one datacenter is attacked. Furthermore, the services in the affected datacenter can be quickly redirected to the secondary datacenter as one of the recovery mechanisms, and vice versa (redirected back to the primary datacenter once service is restored).[1]

Individual workloads include built-in features that manage resource utilization. For example, the throttling mechanisms in Exchange Online and SharePoint Online are part of a multi-layered approach to defending against DoS attacks. Throttling for Exchange Online users is based on the level of resources consumed by the end user, whether the resources are in Active Directory, the Exchange Online information store, or elsewhere. A budget is allocated to each client to limit the resources consumed by a user. Exchange Online throttling for user activity and system components is based on workload management. An Exchange Online workload is an Exchange Online feature, protocol, or service which has been explicitly defined for the purposes of Exchange Online system resource management. Each Exchange Online workload requires system resources such as CPU, mailbox database operations, or Active Directory requests to perform user requests or background work. Examples of Exchange Online workloads include Outlook on the web, Exchange ActiveSync, mailbox migration, and mailbox assistants. Tenant administrators can manage Exchange workload management throttling settings for users with the Exchange Management Shell. There are various forms of throttling which have been implemented within Exchange Online, including PowerShell, Exchange Web Services, and POP and IMAP, Exchange ActiveSync, mobile device connections, recipients, and more. While administrators of on-premises Exchange deployments can configure throttling policies, Administrators cannot configure throttling policies for Exchange Online.

The most common trigger for throttling in SharePoint Online is client-side object model (CSOM) code that performs too many actions at too high a frequency. With CSOM, many actions can be performed with a single request, which can exceed usage limits and cause per-user throttling.

Regardless of the activity which might result in throttling, when a user exceeds usage limits, SharePoint Online throttles any further requests from that user account, usually for a short period of

[1] For more information, see Data Resiliency in Microsoft Office 365.

time. While a user throttle is in effect, all actions by that user are throttled until the throttle expires, according to the following criteria:

- For requests performed by the user directly in a browser, SharePoint Online redirects to a throttling information page, and the requests fail.
- For all other requests, including CSOM calls, SharePoint Online returns HTTP status code 429 ("too many requests"), and the requests fail.

If the offending process continues to exceed usage limits, SharePoint Online may completely block the process and return HTTP status code 503 ("service unavailable").

Vulnerability and Penetration Testing

Microsoft uses many security technologies and practices to protect its cloud infrastructure and on-premises networks against modern, sophisticated threats, including using antimalware components and services for cloud services, virtual machines (VMs), and other systems; Advanced Threat Analytics, which monitors normal usage patterns for networks, systems, and users, and employs machine learning to flag any behavior that is out of the ordinary, and regular penetration testing.

In addition to performing our own penetration tests and offering to our customers a Microsoft Cloud Unified Penetration Testing program, we also engage with third-party security professionals who perform regular vulnerability assessments of and penetration testing against our cloud services. We make the reports from these third-party vulnerability assessments available for download from the Service Trust Preview and the Service Assurance portals.

Summary

The Microsoft Cloud was intentionally built to support a very high volume of activity and to mitigate and protect against application-level DoS attacks through the implementation of throttling, a scaled-out architecture, regional isolation, and high-performance components. To protect the Microsoft Cloud against network-level DoS attacks, we use specialized hardware and application-level DoS protection mechanisms built into our cloud services, as well as network and transport layer DoS protections through an internal Azure-based DoS protection solution.

Ultimately, we realize that we will always be under attack, and that we will never be able to block all attacks. We accept that DoS attacks are part of being in business with online services, and we continue to invest in research and in detection and mitigation strategies. Given our unique characteristics, we also use additional strategies beyond the typical detection and mitigation strategies used in many large enterprises, and instead we employ a strategy that is based on absorption before detection and mitigation. We use defense-in-depth security principles to protect its cloud services and networks. The cornerstone of our strategy implementation is leveraging of our global presence that allows the service to be distributed. Having such a large presence enables us to deflect attacks across a vast surface area.