Defending Office 365 Against Denial-of-Service Attacks

Published: May 24, 2016

Introduction

As a global organization with a significant Internet presence and several prominent Internet properties, Microsoft is a large and common target for hackers and other malicious individuals. In fact, for the last several years, Microsoft has been continuously and persistently under some form of cyber-attack. At just about any given time, at least one of Microsoft's Internet properties is experiencing some form of denial of service traffic. Without reliable and persistent mitigation systems that can defend against these attacks, Microsoft's services would be offline.

This document talks generally about different types of attacks and how Microsoft defends Office 365 and its network against those attacks. Office 365 uses defense-in-depth security principles to protect against internal and external risks. The network--the communication layer between clients and Office 365--is one of the biggest targets of malicious attacks.

Definition and Symptoms of Denial of Service Attacks

One way to attack network services is to create a large number of requests against a service's hosts in an attempt to overwhelm the network and servers in order to deny services to legitimate users. This is referred to as a denial-of-service (DoS) attack. When the attack is performed by multiple actors and vectors, it is referred to as a distributed denial-of-service (DDoS) attack.

Although the means, motives, and targets may 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 particular Web site
- Inability to access a particular 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 DoS Attacks

DoS attacks 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.

Packets/sec Attacks

This attack exploits the simple fact that not all packets are created equal. It focuses on using an excessive numbers 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). As a result of Microsoft's Digital Crimes Unit's work 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 in order 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 DoS Defense

There are three core principles when defending against DoS attacks:

1. Absorption
2. Detection
3. 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 often not economically feasible for organizations to purchase the excess capacity necessary to absorb DoS attacks. So there has to be a balance between absorption, detection, and mitigation. To find that balance, you must understand the attack's growth rate so that you can estimate how much you need to absorb.

Detection is a cat-and-mouse game. You have to 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

In order to successfully defend against a DoS attack, early detection is essential. By detecting an attack before the system is overwhelmed, defenders have the opportunity to execute a response plan.

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

Maximum Capacity / (Maximum Capacity X Growth Rate) = 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, provided that mitigation strategies are used.

Thus, there are really 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 DoS 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 follow. 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. As a result 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.

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 cloud-scale 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, the Office 365 workloads are configured with optimized thresholds based on their protocol and bandwidth usage needs to uniquely protect that workload.

Application-Level Defenses to DoS Attacks

Office 365 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 of 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. The distribution of data in 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).

The throttling mechanisms in Exchange Online and SharePoint Online are also very important in 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 particular 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 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").

Defending Azure against DoS Attacks

Like Office 365, Microsoft Azure is also subject to the threat of attack from multiple sources. To mitigate and protect against the various DoS threats, a highly-scalable and dynamic 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 Azure customers. The Azure DoS mitigation system protects inbound, outbound, and region-to-region traffic.

Most DoS attacks launched against Azure target communications 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 in order 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 VMs supporting an application.

To guard against L3 and L4 attacks the Azure team 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 the Azure network monitoring service to detect attacks. When an attack is detected, automated defense mechanisms kick in.

Summary

Office 365 services are intentionally built to support a very high load 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 Office 365, Microsoft uses application-level DoS protection mechanisms built into Office 365, as well as network and transport layer DoS protections through an internal Microsoft Azure-based DoS protection solution.

Ultimately, Microsoft realizes 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. Microsoft continues to invest in research and in detection and mitigation strategies. Given our unique characteristics, Microsoft uses additional strategies beyond the typical detection and mitigation strategies used in many large enterprises, and instead employs a strategy that is based on absorption before detection and mitigation.

The cornerstone of Microsoft's strategy implementation is leveraging of our global presence that allows the service to be distributed. Microsoft has a significant Internet presence that doubles in size approximately every 18 months. Having such a large presence enables us to deflect attacks across a vast surface area.

Other Materials in this Library

Microsoft publishes a variety of content for customers, partners, auditors, and regulators around security, compliance, risk, trust, privacy, and related areas. Below are links to other content currently in our library. Many of these links point to content available for download from the Microsoft Cloud Service Trust Portal (STP). For information on how to access the STP, see Get started with the Service Trust Portal for Office 365 for business, Azure, and Dynamics CRM Online subscriptions.

Name	Abstract
Auditing and Reporting in Office 365	Describes the auditing and reporting features in Office 365 and Azure Active Directory available to customers. Also details the various audit data that is available to customers via the Office 365 Compliance Center and the Management Activity REST API. Also describes the internal logging data that is available to Microsoft Office 365 engineers for detection, analysis, and troubleshooting.
Data Encryption Technologies in Office 365	Provides an overview of the various encryption technologies that are currently available or recently announced for Office 365, including features deployed and managed by Microsoft, and features managed by customers.
Data Resiliency in Office 365	Describes how Microsoft prevents customer data from becoming lost or corrupt in Exchange Online, SharePoint Online, and Skype for Business, and how Office 365 protects customer data from malware and ransomware.
Defending Office 365 Against Distributed Denial of Service Attacks	Discusses different types of Distributed Denial of Service "DDoS" attacks and how Microsoft defends Office 365 and its network against attacks.
Financial Services Compliance in Microsoft's Cloud Services	Describes how the core contract amendments and the Microsoft Regulatory Compliance Program work together to support financial services customers in meeting their regulatory obligations as they relate to the use of cloud services.
Microsoft Response to New FISC Guidelines in Japan (English) (Japanese)	Explains how Microsoft addresses the risks and requirements described in the FISC Revised Guidelines, and it describes features, controls, and contractual commitments that customers can use to meet the requirements in the Revised Guidelines.
Microsoft Threat, Vulnerability, and Risk Assessment of Datacenter Physical Security	Provides an overview regarding the risk assessment of Microsoft datacenters, including potential threats, controls and processes to mitigate threats, and indicated residual risks.
Office 365 Customer Security Considerations	Provides organizations with quick access to the security and compliance features in Office 365 and considerations for using them.
Office 365 End of Year Security Report 2014	Covers security and legal enhancements made to Office 365 in calendar year 2014 than enables customers and partners to meet legal requirements surrounding independent verification and audits of Office 365.
Office 365 End of Year Security Report and Pen Test Summary 2015	Office 365 End of Year Security Report and Pen Test Summary for CY 2015.
Office 365 Mapping of CSA Cloud Control Matrix 3.0.1	Provides a detailed overview of how Office 365 maps to the security, privacy, compliance, and risk management controls defined in version 3.0.1-11-24-2015 of the Cloud Security Alliance's Cloud Control Matrix.
Office 365 Risk Management Lifecycle	Provides an overview of how Office 365 identifies, evaluates, and manages identified risks.
Office 365 Security Incident Management	Describes how Microsoft handles security incidents in Microsoft Office 365.
Self-Service Handling of Data Spills in Office 365 (restricted to Federal customers via the STP)	Reviews the spillage support provided by Office 365, the tools available to customers, and the configuration settings that should be reviewed in environments that are prone to data spills.
Tenant Isolation in Office 365	Describes how Microsoft implements logical isolation of tenant data within Office 365 environment.