DENIAL OF SERVICE ATTACK MITIGATION ON AWS *"How do I minimize the impact of a DDoS attack on my application?"*

Overview

A Distributed Denial of Service (DDoS) attack is a malicious attempt to make a targeted system, such as a website or application, unavailable to end users. To achieve this, attackers use a variety of techniques that consume network or other resources, interrupting access for legitimate end users.

AWS provides flexible infrastructure and services that help customers implement strong DDoS mitigations and create highly available application architectures that follow AWS Best Practices for DDoS Resiliency. These include services such as Amazon Route 53, Amazon CloudFront, Elastic Load Balancing, and AWS WAF to control and absorb traffic, and deflect unwanted requests. These services integrate with AWS Shield, a managed DDoS protection service that provides always-on detection and automatic inline mitigations to safeguard web applications running on AWS. This document describes common DDoS attack types and provides AWS customers with best practices and strategies for protecting applications from a DDoS attack.

General Best Practices[1]

At its core, DDoS protection and mitigation involves establishing a secure perimeter around your infrastructure and allowing or denying certain traffic based on filters or rules. AWS customers can take advantage of the flexible nature of the cloud and adapt their infrastructure defensively in the event of an attack. When thinking about DDoS security, keep the following best practices in mind:

- **Be ready to scale.** Building infrastructure for scale is fundamental to a well-architected system, however it is also an effective DDoS mitigation technique. Scaling to meet the additional traffic volumes, whether valid or from a DDoS attack, will increase your application's ability to keep running. AWS Shield Advanced includes some cost protection against Amazon EC2, Elastic Load Balancing, Amazon CloudFront, and Amazon Route 53 usage spikes that could result from scaling during a DDoS attack.
- **Minimize the attack surface area and safeguard exposed resources.** In other words: decouple your infrastructure. For example, when running public websites, separate the application from the database and, if possible, the media and static content as well. Decoupled applications limit internet access to critical system components, protecting them from an attack and enabling teams to focus DDoS mitigation efforts on resources that are publicly accessible.
- **Know what is normal; alert on what is not.** In order to get the best help from AWS Support, it is critical that you identify traffic correctly. For example, if a website gets media attention and is suddenly overwhelmed with traffic, blocking that traffic can cause more harm than good. Constant logging and monitoring of infrastructure can help you quickly identify a legitimate attack and engage AWS. For extended visibility into attacks on your Amazon EC2, CloudFront, and Elastic Load Balancing resources, use AWS Shield Advanced to gain exclusive access to advanced, real-time Amazon CloudWatch metrics and reports.
- **Architect for resilience.** Understand the AWS shared responsibility model and take advantage of services and features that can help protect your application. AWS Shield Standard, when used with Amazon CloudFront and Amazon Route 53, provides comprehensive protection against all known infrastructure layer (layer 3 and layer 4) attacks. For additional protection against application layer (layer 7) attacks, use AWS WAF to apply custom mitigation rules.

[1] Adapted from the whitepaper AWS Best Practices for DDoS Resiliency.

Application on AWS

The following sections describe key AWS services involved in DDoS attack mitigation and outline mitigation techniques for common application types.

AWS Services for DDoS Attack Mitigation

AWS offers globally distributed, high network bandwidth and resilient services that, when used in conjunction with application-specific strategies, are key to mitigating DDoS attacks. For more information on how to leverage each of these services and details on how their various features help protect against DDoS attacks, see the whitepaper AWS Best Practices for DDoS Resiliency.

AWS Shield

AWS Shield is a managed DDoS protection service that is available in two tiers: Standard and Advanced. AWS Shield Standard applies always-on detection and inline mitigation techniques, such as deterministic packet filtering and priority-based traffic shaping, to minimize application downtime and latency. AWS Shield Standard is included automatically and transparently to your Elastic Load Balancing load balancers, Amazon CloudFront distributions, and Amazon Route 53 resources at no additional cost. When you use these services that include AWS Shield Standard, you receive comprehensive availability protection against all known infrastructure layer attacks. Customers who have the technical expertise to manage their own monitoring and mitigation of application layer attacks can use AWS Shield together with AWS WAF rules to create a comprehensive DDoS attack mitigation strategy.

AWS Shield Advanced provides enhanced DDoS attack detection and monitoring for application-layer traffic to your Elastic Load Balancing load balancers, CloudFront distributions, Amazon Route 53 hosted zones and resources attached to an Elastic IP address, such Amazon EC2 instances. AWS Shield Advanced uses additional techniques to provide granular detection of DDoS attacks, such as resource-specific traffic monitoring to detect HTTP floods or DNS query floods. AWS Shield Advanced includes 24x7 access to the AWS DDoS Response Team (DRT), support experts who apply manual mitigations for more complex and sophisticated DDoS attacks, directly create or update AWS WAF rules, and can recommend improvements to your AWS architectures. AWS WAF is included at no additional cost for resources that you protect with AWS Shield Advanced.

AWS Shield Advanced includes access to near real-time metrics and reports, for extensive visibility into infrastructure layer and application layer DDoS attacks. You can combine AWS Shield Advanced metrics with additional, fine-tuned AWS WAF metrics for a more comprehensive CloudWatch monitoring and alarming strategy. Customers subscribed to AWS Shield Advanced can also apply for a credit for charges that result from scaling during a DDoS attack on protected Amazon EC2, Amazon CloudFront, Elastic Load Balancing, or Amazon Route 53 resources. See the AWS Shield Developer Guide for a detailed comparison of the two AWS Shield offerings.

AWS WAF

AWS WAF is a web application firewall that helps protect web applications from common web exploits that could affect application availability, compromise security, or consume excessive resources. You can use AWS WAF to define customizable web security rules that control which traffic accesses your web applications. If you use AWS Shield Advanced, you can use AWS WAF at no extra cost for those protected resources and can engage the DRT to create WAF rules.

AWS WAF rules use conditions to target specific requests and trigger an action, allowing you to identify and block common DDoS request patterns and effectively mitigate a DDoS attack. These include size constraint conditions to block a web request based on the length of its query string or request body, and geographic match conditions to implement geo restriction (also known as *geoblocking*) on requests that originate from specific countries. For a complete list of conditions, see the AWS WAF Developer Guide. With AWS WAF, you can also create rate-based rules that automatically block requests from a single IP address if they exceed a customer-defined rate limit. One benefit of rate-based rules is that you can block requests from an IP address while it exceeds the threshold, and then automatically allow requests from that same client once they drop to an acceptable rate. This helps ensure that regular viewers are not held in a persistent block list. You can also combine the rate limit with conditions to trigger different actions for distinct scenarios.

Amazon Route 53

One of the most common targets of DDoS attacks is the Domain Name System (DNS). Amazon Route 53 is a highly available and scalable DNS service designed to route end users to infrastructure running inside or outside of AWS. Route 53 makes it possible to manage traffic globally through a variety of routing types, and provides out-of-the-box shuffle sharding and Anycast routing capabilities to protect domain names from DNS-based DDoS attacks.

Amazon CloudFront

Amazon CloudFront distributes traffic across multiple edge locations and filters requests to ensure that only valid HTTP(S) requests will be forwarded to backend hosts. CloudFront also supports geoblocking, which you can use to prevent requests from particular geographic locations from being served.

Elastic Load Balancing

Elastic Load Balancing automatically distributes incoming application traffic across multiple targets, such as Amazon Elastic Compute Cloud (Amazon EC2) instances, containers, and IP addresses, and multiple Availability Zones, which minimizes the risk of overloading a single resource. Elastic Load Balancing, like CloudFront, only supports valid TCP requests, so DDoS attacks such as UDP and SYN floods are not able to reach EC2 instances. It also offers a single point of management and can serve as a line of defense between the internet and your backend, private EC2 instances. Elastic Load Balancing includes the Application Load Balancer, which is best suited for load balancing of HTTP and HTTPS traffic and also directly supports AWS WAF.

VPCs and Security Groups

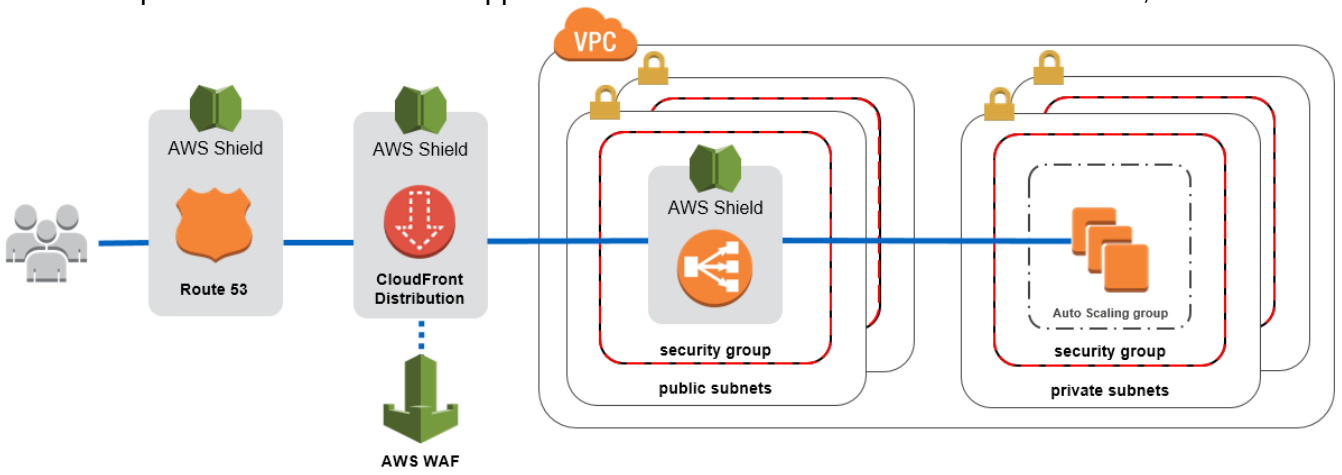
Amazon Virtual Private Cloud (Amazon VPC) allows customers to configure subnet routes, public IP addresses, security groups, and network access control lists in order to minimize application attack surfaces. You can configure load balancers and EC2 instance security groups to allow traffic that originates from specific IP addresses only, such as that from CloudFront or AWS WAF, protecting backend application components from a direct attack.

Mitigation Approaches

In general, there are three major application patterns that we see from customers: *web applications*, *non-web: load balanceable applications*, and *non-web: non-load balanceable applications*. The following sections describe high-level DDoS attack mitigation approaches for each of these application patterns, and include reference architectures that portray the high-level configuration of related services. For detailed information, please see the AWS Best Practices for DDoS Resiliency whitepaper. These architectures assume a highly available design that uses subnets in multiple Availability Zones. They also incorporate AWS Shield Standard services that help manage incoming traffic and minimize the effects of common DDoS attacks.

Web Applications

The reference architecture below represents a stateless web application that relies on HTTP/S for communication, such as a website, web-based API, or mobile application.

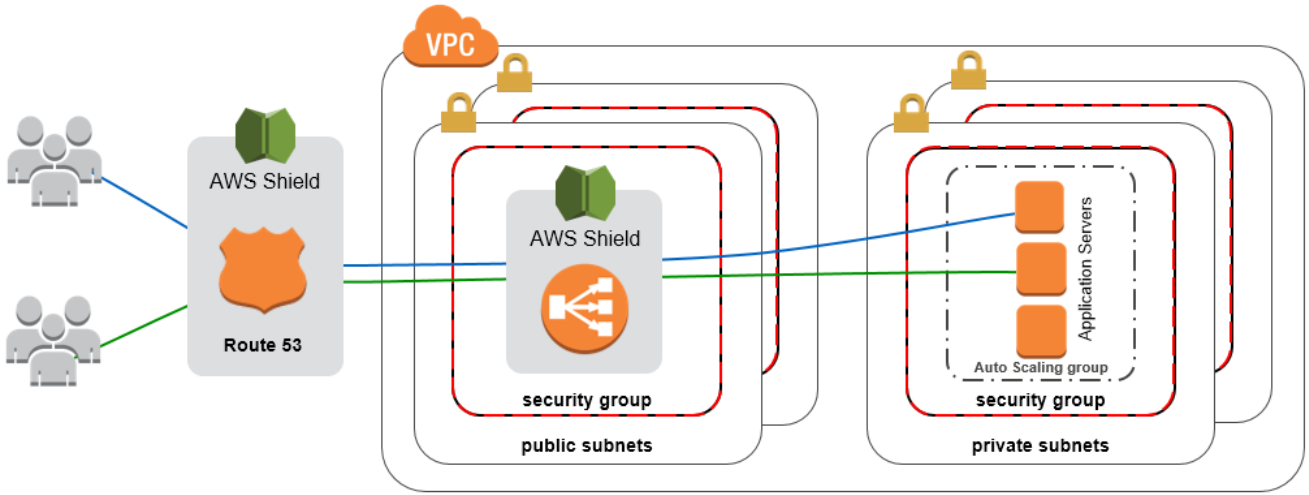


This approach uses Route 53, AWS WAF, CloudFront, and Elastic Load Balancing to control and distribute traffic. Security groups or origin access identity (OAI) can also help minimize the attack surface of backend load balancers, EC2 instances, or Amazon Simple Storage Service (Amazon S3) buckets because they require attackers to make requests through AWS WAF and CloudFront rather than directly from the website origin.

AWS has developed preconfigured AWS WAF rules and tutorials that use AWS Lambda and AWS CloudFormation to help you quickly get started protecting your web applications. AWS Marketplace Sellers offer Managed Rules for AWS WAF, which are a set of pre-configured rules that allow you to quickly get started with AWS WAF rules for your application. AWS WAF Managed Rules are written by security experts who have extensive and up-to-date knowledge of threats and vulnerabilities, and are automatically updated by AWS Sellers as new vulnerabilities and bad actors emerge.

Non-Web: Load Balanceable Applications

The reference architecture below represents a client-server application that requires a TCP connection and general host or session affinity, such as an application using the WebSocket protocol. In this case, the client and the server have a stateful relationship, and the application requires that the client communicates with the same server during the session.

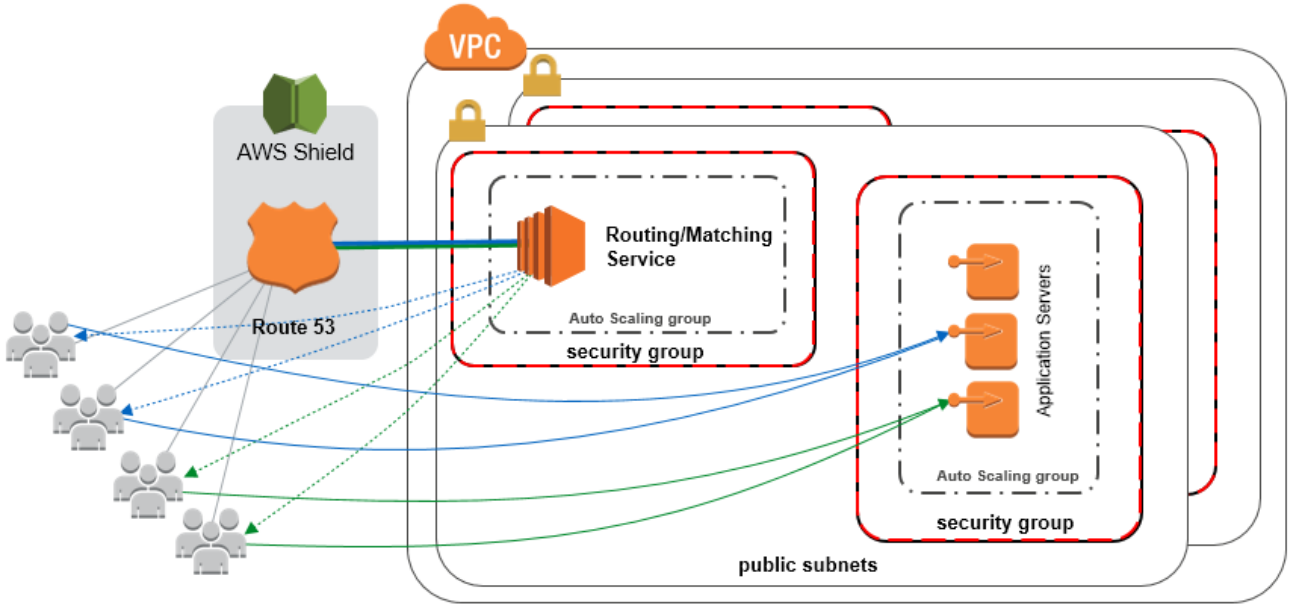


This approach uses Amazon Route 53 and Elastic Load Balancing to control and distribute traffic. As in the previous scenario, you can implement firewall and intrusion-detection products from AWS Marketplace to provide additional traffic monitoring and filtering capabilities to help detect and filter invalid requests.

Another method to mitigate DDoS attacks against TCP-based applications is to vertically scale EC2 instances and use instances that support enhanced networking in order to absorb more connections and traffic. Additionally, host-based IDS/IPS agents can validate the incoming traffic against a rule set and make a decision to serve it or not. When combined with monitoring, timely alerting, and host-based agents to isolate offending traffic, larger instances sizes can provide relief to a targeted application until AWS Support is engaged for additional assistance.

Non-Web: Non-Load Balanceable Applications

The following reference architecture represents a TCP- or UDP-based, non-HTTP service or application, such as DNS, FTP, or a gaming application where user-to-host persistence is required for the users to be able to interact with each other in real-time or near real-time.



This approach uses a highly available, scalable routing service to control and distribute traffic from Route 53. This service acts as a queue for users who are ultimately grouped together by specific attributes. The routing application itself should be either a web application or a non-web, load balanceable application and should follow the previously mentioned techniques for DDoS mitigation. You can introduce additional application-connection logic to the architecture to validate incoming users before providing clients with backend server connection details. For example, implement a highly available lobby or routing server, or a horizontally scalable group of instances configured to check validation keys. Associate Elastic IP addresses to your EC2 instances to further decouple your infrastructure and provide the flexibility to isolate inbound traffic in the event of an attack on an individual instance. Additionally, customers who subscribe to AWS Shield Advanced can enable it on Elastic IP addresses to protect internet facing applications running directly on EC2 instances.

Vertical scaling, EC2 instances that support enhanced networking, and host-based IDS/IPS agents can help mitigate a DDoS attack in this scenario as well. For workloads that have a predetermined set of external consumers, configure security groups to allow traffic that originates from specific IP addresses only.

Resources

AWS Shield	https://aws.amazon.com/shield/
AWS Best Practices for DDoS Resiliency	https://d0.awsstatic.com/whitepapers/Security/DDoS_White_Paper.pdf
Elastic Load Balancing	https://aws.amazon.com/elasticloadbalancing/
Amazon Route 53	https://aws.amazon.com/route53/
AWS WAF	https://aws.amazon.com/waf/
Amazon CloudFront	https://aws.amazon.com/cloudfront/
AWS WAF Security Automations	https://aws.amazon.com/answers/security/aws-waf-security-automations/