Best Practices for DDoS Protection and Mitigation on Google Cloud Platform

Last updated: April 12th, 2016

Introduction

A Denial of Service (DoS) attack is an attempt to render your service or application unavailable to your end users. With Distributed Denial of Service (DDoS) attacks, the attackers use multiple resources (often a large number of compromised hosts/instances) to orchestrate large scale attacks against targets. This document describes the best practices for protecting against and mitigating such DDoS attacks for your Google Cloud Platform (GCP) deployment.

Protecting Shared Infrastructure

Google has mechanisms in place to protect its cloud infrastructure and its production services. These mechanisms are designed to ensure that no single service can overwhelm the shared infrastructure and to provide isolation among customers using the shared infrastructure. Details of these mechanisms are out of scope for this document.

DDoS Protection and Mitigation for your GCP Deployment

Successfully thwarting and handling DDoS attacks for your GCP deployment is a shared responsibility between Google Cloud Platform and you. DDoS defense involves deploying detection systems, implementing barriers and being able to absorb attacks by scaling in order to prevent attackers from overwhelming or disabling access to your services or applications. Google Cloud Platform provides several of these mechanisms automatically and you can follow the best practices detailed below on your end to help secure your GCP deployment:

- **Reduce the attack surface for your GCE deployment**
    - Provision your own isolated and secure piece of the Google Cloud with Google Cloud Virtual Network. View the best practice here.

  - Isolate and secure your deployment using subnetworks and networks, firewall rules, tags and Identity and Access Management (IAM).
  - Open access to ports and protocols that you need using firewall rules and/or protocol forwarding.
  - GCP provides anti-spoofing protection for the private network (IP addresses) by default.
  - GCP automatically provides isolation between virtual networks.

- **Isolate your internal traffic from the external world**
  - Deploy instances without public IPs unless necessary.
  - You can set up a NAT gateway or SSH bastion to limit the number of instances that are exposed to the internet.
  - Once available, deploy Internal Load Balancing for your internal client instances accessing internally deployed services thereby avoiding exposure to the external world. [Internal LB expected to be available in the second half of 2016.]

- **DDoS Protection by enabling Proxy-based Load Balancing**
  - When you enable HTTP(S) Load Balancing or SSL proxy Load Balancing, Google infrastructure mitigates and absorbs many Layer 4 and below attacks, such as SYN floods, IP fragment floods, port exhaustion, etc.
  - If you have HTTP(S) Load Balancing with instances in multiple regions, you are able to disperse your attack across instances around the globe.

- **Scale to absorb the attack**
  - **Protection by Google Frontend infrastructure**
    With Google Cloud Global Load Balancing, the frontend infrastructure which terminates user traffic, automatically scales to absorb certain types of attacks (e.g., SYN floods) before they reach your compute instances.
  - **Anycast-based Load Balancing**: HTTP(S) Load Balancing and SSL proxy enable a single anycast IP to front-end your deployed backend instances in all regions. Normally your user traffic is directed to the closest backend with capacity; in the event of a DDoS attack, the additional advantage of this approach is that it increases the surface area to absorb this attack by moving traffic to instances with available capacity in any region where backends are deployed.
  - **Autoscaling**: When you configure HTTP(S)or SSL Proxy Load Balancing, Google frontend infrastructure that terminates your user traffic protects your backends. You should also provision sufficient number of instances

and/or configure autoscaling to handle spikes in traffic. In the event of a sudden traffic spike, the load balancing proxy layer will distribute the traffic across all the backends with available capacity. In parallel, the autoscaler ramps up the backends inline with traffic that needs to be handled.

- **Protection with CDN Offloading**
    - Google Cloud CDN acts as a proxy between your clients and your origin servers. For cacheable content, Cloud CDN caches and services this content from points-of-presence (POPs) closer to your users as opposed to sending them to backend servers (instances). In the event of DDoS attack for cacheable content, the requests are sent to POPs all over the globe as opposed to your origin servers, thereby providing a larger set of locations to absorb the attack.
    - If you use CDN Interconnect, you can leverage the additional DDoS Protection provided by our CDN Interconnect partners. You can check the partner page for specifics on their DDoS protection capabilities.

- **Deploy third-party DDoS protection solutions**
    - In order to meet your specific needs of protection for DDoS attack prevention/mitigation, consider purchasing specialized third-party DDoS protection solutions to protect against such attacks.
    - You can also deploy DDoS solutions available via Google Cloud Launcher.

- **App Engine deployment**
    - App Engine is designed to be a fully multi-tenant system and implements a number of safeguards intended to ensure that a single bad application will not impact the performance or availability of other applications on the platform.
    - App Engine sits behind the Google Front End which mitigates and absorbs many Layer 4 and below attacks, such as SYN floods, IP fragment floods, port exhaustion, etc.
    - You can also specify a set of IPs/IP networks via a dos.yaml file to block them from accessing your application(s).

- **Google Cloud Storage**
    - If you do not want to require your users to have a Google account in order to be able to access your Google Cloud Storage resources, you can control access using signed URLs.

- **API rate-limiting**
  - API rate limits define the number of requests that can be made to the Google Compute Engine API. API rate limits apply on a per-project basis.
  - Currently, projects are limited to an API rate limit of 20 requests/second.

- **Resource Quotas**
  - Compute Engine enforces quotas on resource usage for a variety of reasons. For example, quotas protect the community of Google Cloud Platform users by preventing unforeseen spikes in usage. Special quotas limit access for projects that are just exploring Google Cloud Platform on a free trial basis.

Conclusion

Google Cloud Platform provides a number of features to defend against DDoS attacks. You can use these in conjunction with the above mentioned best practices and other measures tailored to your requirements to make your GCP deployment resilient to DDoS attacks.