the Availability Digest

www.availabilitydigest.com
@availabilitydig

# Cascading Software Bugs Take Down Google Compute Engine

April 2016

On April 11, 2016, Google's Compute Engine (GCE) had a massive outage that affected all of Google's regions worldwide. The outage was caused by a series of software bugs that fed on each other while Google engineers were busy upgrading their network. The first software bug caused the network upgrade to become corrupted. The second software bug sent the corrupted upgrade to the network rather than cancelling it. The third software bug failed to inform the network management software that a corrupted network upgrade was being propagated.

The result was that inbound Internet traffic to Google was not routed correctly. Connections were dropped and users could not reconnect. Services dependent upon the network such as VPNs and Level 3 load balancers began to fail. Google users worldwide were unable to connect to the Google Compute Engine. Outbound Internet traffic was not affected.

The asia-east1 region was unreachable for over an hour. The entire Google worldwide GCE network was down for eighteen minutes.

Google announced service refunds to its clients. They exceeded the requirements of Google's SLAs.

## The Google Compute Engine

The Google Compute Engine is the Infrastructure-as-a-Service (IaaS) component of the Google Cloud Platform. It is fundamental to the global infrastructure that runs Google's search engine, Gmail, You Tube, and other Google services.

The GCE enables users to launch virtual machines (VMs) on demand. VMs can be launched from standard Google images or from custom images created by users. VMs run under the KVM hypervisor and support Windows or Linux guest operating systems.

## Google IP Blocks

Google uses contiguous blocks of internet addresses for users to connect to Google services. It calls these blocks *IP blocks*. IP blocks are announced to the rest of the Internet via the industry-standard Border Gateway Protocol (BGP).[1] This announcement allows systems outside of Google's network to 'find' GCP services regardless of the network to which the systems are connected.

To maximize service performance, Google's networking systems announce the same IP blocks from several different locations in its network around the world. This allows users to take the shortest available path through the Internet to reach their desired Google service.

[1] Eavesdropping on the Internet, *Availability Digest*; March 2009.
http://www.availabilitydigest.com/public_articles/0403/bgp.pdf

This approach also enhances reliability. If a user is unable to reach one location announcing an IP block due to an Internet failure between the user and Google, the user will be sent to the next closest point of announcement.

## The Corrupted Network Upgrade

On Monday, April 11, 2016, Google engineers decided to remove an unused GCE IP block from the network configuration. They instructed the Google systems to propagate the new configuration across the network. This sort of change had been performed many times previously without incident.

***Software Bug 1***

However, on this occasion, the network configuration management software detected an inconsistency in the newly supplied configuration. The detection of the inconsistency was triggered by a timing quirk in the IP block removal procedure. The IP block had been removed from one configuration file, but the removal had not yet propagated to a second configuration file. The network configuration management software deemed this to be a failure in the upgrade.

The network configuration management software is designed to be 'fail safe' and to revert to its current configuration if it detects a problem rather than proceeding with the new configuration.

***Software Bug 2***

However, a second software bug reared its ugly head. Instead of retaining the previously known good configuration, the network configuration management software instead removed all GCE IP blocks from the network configuration. It then began to propagate this new (now empty) configuration across the network.

## The Sick Canary

Google's networking systems have a number of safeguards to prevent the propagation of incorrect configurations. One such safeguard is the 'canary step.'

Rather than immediately deploying the new network configuration across the entire network, it is first deployed to a single site. If it works properly at that site, it is deployed to a few more sites. In this way, a failure hopefully can be detected before it becomes widespread.

***Software Bug 3***

The canary step indeed identified the new configuration as being unsafe. However, another software bug did not send the canary step's conclusion back to the network management software. Therefore, the network management software concluded that the new configuration was safe and began it progressive rollout.

## The Google GCE Network Fails

As the rollout progressed, those sites that had been announcing GCE IP blocks ceased to do so when they received the new configuration with no IP blocks. As more and more sites stopped announcing GCE IP blocks, the network continued to send GCE traffic to the remaining sites that were still announcing GCE IP blocks. However, user communication latency was rising as more and more users were sent to sites that were not close to them and as those sites got hit with ever increasing traffic.

The first area to be noticeably affected was the asia-east1 region. It finally lost connectivity at 18:14 PM U.S. Pacific time. Google engineers had been trying to determine the cause of the asia-east1 problems.

Fifty-three minutes after asia-east1 lost connectivity, at 19:07, the last site announcing GCE IP blocks received the configuration. Now, with no sites announcing IP blocks, internet traffic to the GCE dropped quickly. Two minutes later, at 19:09, it had dropped by 95%.

Because the outage took down all regions, it made it difficult if not impossible for clients to mitigate the impact of the outage. However, the outage did not affect the Google App Engine, Google Cloud Storage, or other Google Cloud Platform products.

## The Aftermath

The Google engineers now knew they had a widespread problem. Without knowing the cause, they immediately backed out the configuration changes and propagated the original configuration throughout the network. This action was successful at ending the outage at 19:27. In total, the entire GCE network had been down for eighteen minutes. The asia-east 1 region had been down for one hour and thirteen minutes.

Having frozen all configuration changes, the Google engineers worked through the night to ensure the systems were stable and to determine the root cause of the problem. By 7 AM the next morning, they were confident that they had established the cause as software bugs in the network configuration management software. They determined that the GCE was not at risk of a reoccurrence of the problem.

The Google engineering teams then turned to the task of identifying a broad array of prevention, detection, and mitigation systems intended to add additional defenses against similar problems in the future. After just the first day, they already had planned fourteen distinct engineering changes spanning prevention, detection, and mitigation.

## Google's Reimbursement

Google is reimbursing all affected Google Compute Engine users with service credits of 10% of the monthly charges for GCE clients and 25% of the monthly charges for VPN clients. These reimbursements exceed Google's SLA requirements.[2]

## Lessons Learned

Google learned several lessons from this outage:

- There must be safeguards to prevent a failure in a progressive rollout from being masked by a failure of the monitoring system.
- They should monitor for a decrease in capacity or redundancy even when the system is functioning properly.
- IP block announcements should be compared before and after a configuration change to ensure that they are still correct.
- Network configurations should be checked to ensure that they contain specific IP blocks.

[2] https://cloud.google.com/compute/sla
https://cloud.google.com/vpn/sla

## Acknowledgements

Thanks to our subscriber, Gary Dick, for pointing us to this topic.

Information for this article was taken from the following resources:

Google Post Mortem, https://status.cloud.google.com/incident/compute/16007?post-mortem; April 13, 2016.
Google Reimburses Cloud Clients After Massive Google Compute Engine Outage, *Talkin' Cloud*; April 14, 2016.
Google Compute Engine, *Wikipedia*.