Root Cause Analysis Report

Google Compute Engine Incident #16015

Problem Statement

Report Number	RCA-2016-19-08-2016-040	**RCA Owner**	Chris Eckert
Report Date		**RCA Facilitator**	Jon Boisoneau

Focal Point: Negative impact to Google Cloud for 99 minutes

When

Start Date: 8/5/2016	End Date: 8/5/2016
Start Time: 00:55 PDT	End Time: 02:34 PDT
Unique Timing	After a router failed and after using a new configuration while changing out the router.

Where

Other	Data center - undisclosed location
Other	Router - undisclosed type

Actual Impact

Cost	Cost of investigation	$0.00
Cost	Cost of repairs	$0.00
Safety	Do safety-critical operations rely on this connectivity?	$0.00
Reputation (External)	Damage to Google's reputation for reliable service	$0.00
Revenue	Possible impact to advertising revenue	$0.00
		Actual Impact Total: $0.00

Frequency	
Frequency Note	Frequency undisclosed

Potential Impact

Potential Impact Total: $0.00

Report Summaries

Executive Summary

No executive summary required - see detailed summary below.

Cause and Effect Summary

NOTE FROM SOLOGIC: This summary was provided by Google. We used this summary to create the cause and effect chart.

SUMMARY:

On Friday 5 August 2016, some Google Cloud Platform customers experienced increased network latency and packet loss to Google Compute Engine (GCE), Cloud VPN, Cloud Router and Cloud SQL, for a duration of 99 minutes. If you were affected by this issue, we apologize. We intend to provide a higher level reliability than this, and we are working to learn from this issue to make that a reality.

DETAILED DESCRIPTION OF IMPACT:

On Friday 5th August 2016 from 00:55 to 02:34 PDT a number of services were disrupted:

Some Google Compute Engine TCP and UDP traffic had elevated latency. Most ICMP, ESP, AH and SCTP traffic inbound from outside the Google network was silently dropped, resulting in existing connections being dropped and new connections timing out on connect.

Most Google Cloud SQL first generation connections from sources external to Google failed with a connection timeout. Cloud SQL second generation connections may have seen higher latency but not failure.

Google Cloud VPN tunnels remained connected, however there was complete packet loss for data through the majority of tunnels. As Cloud Router BGP sessions traverse Cloud VPN, all sessions were dropped.

All other traffic was unaffected, including internal connections between Google services and services provided via HTTP APIs.

ROOT CAUSE:

While removing a faulty router from service, a new procedure for diverting traffic from the router was used. This procedure applied a new configuration that resulted in announcing some Google Cloud Platform IP addresses from a single point of presence in the southwestern US. As these announcements were highly specific they took precedence over the normal routes to Google's network and caused a substantial proportion of traffic for the affected network ranges to be directed to this one point of presence. This misrouting directly caused the additional latency some customers experienced.

Additionally this misconfiguration sent affected traffic to next-generation infrastructure that was undergoing testing. This new infrastructure was not yet configured to handle Cloud Platform traffic and applied an overly-restrictive packet filter. This blocked traffic on the affected IP addresses that was routed through the affected point of presence

to Cloud VPN, Cloud Router, Cloud SQL first generation and GCE on protocols other than TCP and UDP.

REMEDIATION AND PREVENTION:

Mitigation began at 02:04 PDT when Google engineers reverted the network infrastructure change that caused this issue, and all traffic routing was back to normal by 02:34. The system involved was made safe against recurrences by fixing the erroneous configuration. This includes changes to BGP filtering to prevent this class of incorrect announcements.

We are implementing additional integration tests for our routing policies to ensure configuration changes behave as expected before being deployed to production. Furthermore, we are improving our production telemetry external to the Google network to better detect peering issues that slip past our tests.

Solutions

SO-0001	**Solution**	Implement additional integration tests for routing policies.		
	Cause(s)	Replacement procedure applied new configuration		
	Note	We are implementing additional integration tests for our routing policies to ensure configuration changes behave as expected before being deployed to production.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	9/2/2016	**Status**	Completed
	Term	short	**Cost**	
SO-0002	**Solution**	Revert the network infrastructure changes.		
	Cause(s)	Router was replaced		
	Note	Mitigation began at 02:04 PDT when Google engineers reverted the network infrastructure change that caused this issue, and all traffic routing was back to normal by 02:34. The system involved was made safe against recurrences by fixing the erroneous configuration. This includes changes to BGP filtering to prevent this class of incorrect announcements.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	9/2/2016	**Status**	Completed
	Term	short	**Cost**	
SO-0003	**Solution**	Improve production telemetry external to the Google network to better detect peering issues that slip past tests.		
	Cause(s)	Replacement procedure applied new configuration		
	Note	No additional notes provided.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	9/2/2016	**Status**	Completed
	Term		**Cost**	

Team

Facillitator	**Owner**
Jon Boisoneau	Chris Eckert
jon.boisoneau@sologic.com	chris.eckert@sologic.com

Participants

Brian Hughes	Cory Boisoneau	Chris Eckert
brian.hughes@sologic.com	cory.boisoneau@sologic.com	chris.eckert@sologic.com

Evidence

EV-0001	**Evidence**	Google Cloud Status Dashboard
	Cause(s)	99 Minutes to Recover Causes of event Certain requests were highly specific Google Cloud SQL 1st Gen connections dropped Google Cloud SQL 2nd Gen connections had higher latency, but not dropped Google Cloud VPN tunnels experienced complete packet loss Google Compute Engine TCP, UDP traffic = Elevated latency Highly specific requests take precedence ICMP, ESP, AH, SCTP inbound traffic silently dropped Large amount of traffic sent to single point of presence Negative customer impacts (list) New connections timed out upon connect New hardware not configured to handle Cloud traffic New hardware undergoing testing Next gen infrastructure had overly restrictive packet filter No redundancy? Other traffic unaffected Replacement procedure applied new configuration Router was faulty Router was replaced Single point of presence could not process all the traffic Testing requires restrictive configuration Traffic sent to next generation infrastructure
	Location(s)	https://status.cloud.google.com/incident/compute/16015
	Attachment(s)	
	Contributor	
	Type	Other
	Quality	★★★☆☆

Actions

AC-0001	**Action**	Find out why other traffic was unaffected. Could this help us understand how to protect the impacted area?
	Cause(s)	Other traffic unaffected
	Assigned	
	Date	8/22/2016
AC-0002	**Action**	Find out if this recovery time duration is abnormally long, or if it was acceptable given the scenario. The purpose of finding out this info is to identify more effective ways of responding to outages in the future.
	Cause(s)	99 Minutes to Recover
	Assigned	
	Date	8/22/2016
AC-0003	**Action**	Why weren't these 2nd Gen connections dropped?
	Cause(s)	Google Cloud SQL 2nd Gen connections had higher latency, but not dropped
	Assigned	
	Date	8/22/2016
AC-0004	**Action**	Why would the specificity of requests matter?
	Cause(s)	Certain requests were highly specific
	Assigned	
	Date	8/22/2016
AC-0005	**Action**	Why do specific requests take precedence?
	Cause(s)	Highly specific requests take precedence
	Assigned	
	Date	8/22/2016
AC-0006	**Action**	Why was the router faulty? Is there a history of this type of failure? How many routers of this type do we operate? Is there an abnormal risk of additional failures?
	Cause(s)	Router was faulty
	Assigned	
	Date	8/22/2016

AC-0007	**Action**	Was this change tested? Was a failure of this type predicted? Why not? Could this result have been anticipated? How can we make adjustments to our change management procedures to help trap this kind of situation in the future?
	Cause(s)	Replacement procedure applied new configuration
	Assigned	
	Date	8/22/2016
AC-0008	**Action**	If so many critical processes depend on this router, why is there no redundancy?
	Cause(s)	No redundancy?
	Assigned	
	Date	8/22/2016
AC-0009	**Action**	We need to know a bit more about this next-generation infrastructure and how sending traffic to it impacted this issue.
	Cause(s)	Traffic sent to next generation infrastructure
	Assigned	
	Date	8/22/2016

Notes

NO-0001	**Note**	Note - this cause acts as a single landing point for the connecting branches.
	Cause(s)	Causes of event

Chart Key

- Transitory
- Non Transitory
- Transitory Omission
- Non Transitory Omission
- Undefined
- Chart Quality Alert
- Focal Point
- Evidence
- Notes
- Solutions
- Actions

Negative impact to Google Cloud for 99 minutes

- Negative customer impacts (list)
  - AND OR
    - Google Compute Engine TCP, UDP traffic = Elevated latency
      - c
      - d
      - b
      - a
      - Causes of event
        - Large amount of traffic sent to single point of presence
          - Router was replaced
            - Solutions
              - Revert the network infrastructure changes.
              - Criteria: Pass — Status: Completed
              - Mitigation began at 02:04 PDT when Google engineers reverted the network infrastructure change that caused this issue, and all traffic routing was back to normal by 02:34. The system involved was made safe against recurrences by fixing the erroneous configuration. This includes changes to BGP filtering to prevent this class of incorrect announcements.
            - Router was faulty
            - No redundancy?
          - Replacement procedure applied new configuration
            - Solutions
              - Implement additional integration tests for routing policies.
              - Criteria: Pass — Status: Completed
              - We are implementing additional integration tests for our routing policies to ensure configuration changes behave as expected before being deployed to production.
              - Improve production telemetry external to the Google network to better detect peering issues that slip past tests.
              - Criteria: Pass — Status: Completed
              - No additional notes provided.
          - Certain requests were highly specific
            - Highly specific requests take precedence
          - Single point of presence could not process all the traffic
            - Traffic sent to next generation infrastructure
            - Next gen infrastructure had overly restrictive packet filter
              - New hardware not configured to handle Cloud traffic
                - New hardware undergoing testing → END — Terminated Because: Desired State
                - Testing requires restrictive configuration → END — Terminated Because: Other causual paths more productive
    - ICMP, ESP, AH, SCTP inbound traffic silently dropped — Connects To: a Causes of event
    - New connections timed out upon connect — Connects To: b Causes of event
    - Google Cloud SQL 1st Gen connections dropped — Connects To: c Causes of event
    - Google Cloud VPN tunnels experienced complete packet loss — Connects To: d Causes of event
    - Google Cloud SQL 2nd Gen connections had higher latency, but not dropped
- Other traffic unaffected
- 99 Minutes to Recover