Google Cloud Status Dashboard

This page provides status information on the services that are part of Google Cloud Platform. Check back here to view the current status of the services listed below. If you are experiencing an issue not listed here, please contact Support. Learn more about what's posted on the dashboard in this FAQ. For additional information on these services, please visit cloud.google.com.

Google App Engine Incident #17005

High Latency in App Engine

Incident began at **2017-06-07 14:21** and ended at **2017-06-07 15:30** (all times are **US/Pacific**).

DATE	TIME	DESCRIPTION
Jun 13, 2017	10:07	ISSUE SUMMARY On Wednesday 7 June 2017, Google App Engine experienced highly elevated serving latency and timeouts for a duration of 138 minutes. If your service or application was affected the increase in latency, we sincerely apologize – this is not the level of reliability and performance we expect of our platform, and we are taking immediate steps to improve the platform's performance and availability. DETAILED DESCRIPTION OF IMPACT On Wednesday 7 June 2017, from 13:34 PDT to 15:52 PDT, 7.7% of active applications on the Google App Engine service experienced severely elevated latency; requests that typically take under 500ms to serve were taking many minutes. This elevated latency would have either resulted in users seeing additional latency when waiting for responses from the affected applications or 500 errors if the application handlers timed out. The individual application logs would have shown this increased latency or increases in "Request was aborted after waiting too long to attempt to service your request" error messages. ROOT CAUSE The incident was triggered by an increase in memory usage across all App Engine appservers in a datacenter in us-central. An App Engine appserver is responsible for creating instances to service requests for App Engine applications. When its memory usage increases to unsustainable levels, it will stop some of its current instances, so that they can be rescheduled on other appservers in order to balance out the memory requirements across the datacenter. This transfer of an App Engine instance between appservers consumes CPU resources, a signal used by the master scheduler of the datacenter to detect when it must further rebalance traffic across more appservers (such as when traffic to the datacenter increases and more App Engine instances are required). Normally, these memory management techniques are transparent to customers but in isolated cases, they can be exacerbated by large amounts of additional traffic being routed to the datacenter, which requires more instances to service user requests. The increased load and memory requirement from scheduling new instances combined with rescheduling instances from appservers with high memory usage resulted in most appservers being considered "busy" by the master scheduler. User requests needed to wait for an available instance to either be transferred or created before they were able to be serviced, which results in the increased latency seen at the app level. REMEDIATION AND PREVENTION Latencies began to increase at 13:34 PDT and Google engineers were alerted to the increase in latency at 13:45 PDT and were able to identify a subset of traffic that was causing the increase in memory usage. At 14:08, they were able to limit this subset of traffic to an isolated partition of the datacenter to ease the memory pressure on the remaining appservers. Latency for new requests started to improve as soon as this traffic was isolated; however, tail latency was still elevated due to the large backlog of requests that had accumulated since the incident started. This backlog was eventually cleared by 15:52 PDT. To prevent further recurrence, traffic to the affected datacenter was rebalanced with another datacenter. To prevent future recurrence of this issue, Google engineers will be re-evaluating the resource distribution in the us-central datacenters where App Engine instances are hosted. Additionally, engineers will be developing stronger alerting thresholds based on memory pressure signals so that traffic can be redirected before latency increases. And finally, engineers will be evaluating changes to the scheduling strategy used by the master scheduler responsible for scheduling appserver work to prevent this situation in the future.
Jun 07, 2017	16:02	The issue with Google App Engine displaying elevated error rate has been resolved for all affected projects as of 15:30 US/Pacific. We will conduct an internal investigation of this issue and make appropriate improvements to our systems to help prevent or minimize future recurrence. We will provide a more detailed analysis of this incident once we have completed our internal investigation.
Jun 07, 2017	14:52	The issue with Google App Engine displaying elevated error rate should be resolved for the majority of projects and we expect a full resolution in the near future. We will provide another status update by 15:30 US/Pacific with current details.
Jun 07, 2017	14:21	We have identified an issue with App Engine that is causing increased latency to a portion of applications in the US Central region. Mitigation is under way. We will provide more information about the issue by 15:00 US/Pacific.

All times are US/Pacific

Send Feedback