Storage Frontend Postmortem; Google incident #123, JaneCorp incident #22

Date: 2017-10-26

Authors: Adrian H, Gwendolyn S, Dave R (Google); Jane D (JaneCorp)

Status: Complete, action items in progress

Summary: New storage frontend release caused increased end-user latency for uncached objects of a common size in a single region.

Background:
Google's storage frontend service ("storage FE") handles all customer read/write requests for binary objects for the Google Cloud Blob storage product. It is deployed on a regional basis, with one pool of frontends in each Cloud region and no cross-region dependencies. It is managed by the Storage SRE team who control configuration pushes and binary releases to the service. The frontends in each region share a cache to reduce the traffic load on the underlying storage system. Over 70% of read requests are cache hits. Object sizes read through storage FE vary from 1KB to 1MB.

The release process for storage FE happens weekly. There is a deployment to a single Cloud region (us-arctic1) on Day 1, before it goes to 2 other regions on Day 2, and all regions by Day 4. This is to mitigate the impact of a bad release. Before release there is an integration testing process that aims to identify significant performance regressions.

Monitoring and alerting of storage FE is broken down by Cloud region, so there is a distinct monitoring console and alerting configuration for us-arctic1. Monitoring includes the cache hit rate and overall read latency, but there is not a breakdown of latency by object size.

Cloud customer JaneCorp uses Cloud Blob as a backend for their JaneApp application. This has a 99.9% availability target and 1000ms @ 95% latency SLO. Most of their requests are cache misses in storage FE, and they use objects in the 4KB-32KB size range. They serve from two regions: us-arctic1 and europe-iceland1.

Impact:

- 3% of all storage FE read requests (20% of cache misses) in us-arctic1 region served with out-of-SLO latency for 2.5 hours.
- Customer app JaneApp served 50% of requests with out-of-SLO latency for 2 hours, burning 1.5x their 30 day error budget.

Root Causes:

- incorrect implementation of a new object lookup function;
- bug only manifested on uncached objects of a particular size;

- qualification tests did not separately measure latency impact by object size and cache status;
- production monitoring did not measure or alert on increased latency by object size or cache status.

Trigger: Binary rollout.

Resolution: Rolled back binary and verified that latency returned to normal. Identified and reverted offending code change.

Detection: Customer monitoring.

Action items:

Action Item	**Type**	**Owner**	**Bug / status**
Extend storage FE integration tests to break down latency by cache status and object size bucket	*prevent*	storage dev (Google)	b-12345 / **complete**
Implement per-region storage FE latency monitoring broken down by cache status and object size bucket.	*detect*	storage SRE (Google)	b-23456 / *in progress*
Update JaneApp Ops playbook with accurate instructions for a regional failover.	*mitigate*	Ops (JaneCorp)	MC-54321 / **complete**
Implement and document thru-stack latency breakdown for JaneApp	*detect*	JaneApp dev (JaneCorp)	MC-54322 / *in progress*
Schedule monthly region failover drills for JaneApp.	mitigate	Ops (JaneCorp)	MC-54323 / *not started*

Lessons learned

What went well:

- Storage FE rollback was quick once the problem was identified.
- The shared monitoring for JaneApp enabled Google Cloud Support to quickly confirm the latency problem and location and escalate.
- The 1-region-per-day rollout schedule limited the impact of the bug to a single region.
- JaneApp monitoring and alerting paged the on-call for high service latency within 10 minutes of the rollout.
- JaneCorp Ops had an existing plan to fail out of a region as mitigation for a single-region Google service problem.

What went badly:

- Storage FE Integration / qualification testing did not detect the problem. It does measure latency of the new binary, but does not break this down by object cache status or size.
- Storage FE per-region production monitoring did not clearly show the problem, similarly due to too-broad aggregation.
- No alert fired for the affected region.
- Customer detected the problem (rather than Google).
- JaneApp monitoring did not show thru-stack latency breakdown, making it hard to identify the source of the latency.
- JaneApp was out of latency SLO for 50% of their traffic for 2 hours, blowing their 30 day error budget.
- JaneApp took 30 minutes to fail out of the affected region due to stale instructions in the Ops playbook.

Where we got lucky:

- Only reads of objects between 4KB and 32KB in size were materially affected, and only uncached objects.
- JaneApp was sufficiently affected in the one region to detect the problem and alert Google before it went to other regions.
- The incident happened during JaneCorp business hours, when developers were on hand to help debug.

Timeline

All times Pacific.

2017-10-19

10:14 Storage FE developer commits change c-12345 improving object lookup reliability. This contains a bug that makes reads of objects of size 4KB-32KB take 3x as long as normal to retrieve.

17:00 Change c-12345 incorporated into release branch r-777

2017-10-23

13:15 Integration testing of r-777 completes. Overall read latency is increased by about 1% but this is not perceptible.

2017-10-26

08:00 Storage FE oncall triggers the rollout of release 555 of Storage FE to the Day 1 region (us-arctic1)

08:10 us-arctic1 storage FE pool starts serving with the new release **<OUTAGE BEGINS>**

08:20 us-arctic1 overall storage FE read latency increases by 2% from baseline [graph]

08:25 JaneApp pages oncall for out-of-SLO latency in us-arctic1 region **<CUSTOMER DETECTION>**

08:40 JaneApp oncall starts diagnosis work

09:10 JaneApp oncall can't identify latency source from available monitoring, contacts JaneApp developer

09:40 JaneApp developer and oncall confirm that the latency is coming from Cloud Blob service.

09:42 JaneApp oncall announces intention to fail JaneApp out of us-arctic1, starts

following the playbook instructions

09:50 Playbook instructions aren't working, traffic is still flowing to us-arctic1

10:05 JaneApp secondary oncall raises P1 Google Support ticket #67890: "High latency for Cloud Blob in us-arctic1 region"

10:10 JaneApp successfully failed out of us-arctic1 **<CUSTOMER MITIGATION>**

10:18 Google Cloud Support engineer confirms on ticket that they see the problem in the JaneApp shared monitoring, pages storage FE oncall

10:25 storage FE oncall acks page, starts investigation

10:35 storage FE oncall can't see any clear latency regression in us-arctic1, starts digging into logs.

10:45 storage FE oncall identifies that uncached objects might have some latency regression. [graph]

10:50 storage FE oncall has sufficient suspicion of the recent rollout and starts a rollback of release 555 in us-arctic1.

11:10 us-arctic1 storage FE pool now serving with original release **<OUTAGE ENDS>**

13:30 Change c-12345 identified as suspicious.

14:15 Integration testing re-run with object size breakdown, confirms regression.

14:30 Change c-12345 rolled back. **<RESOLUTION>**