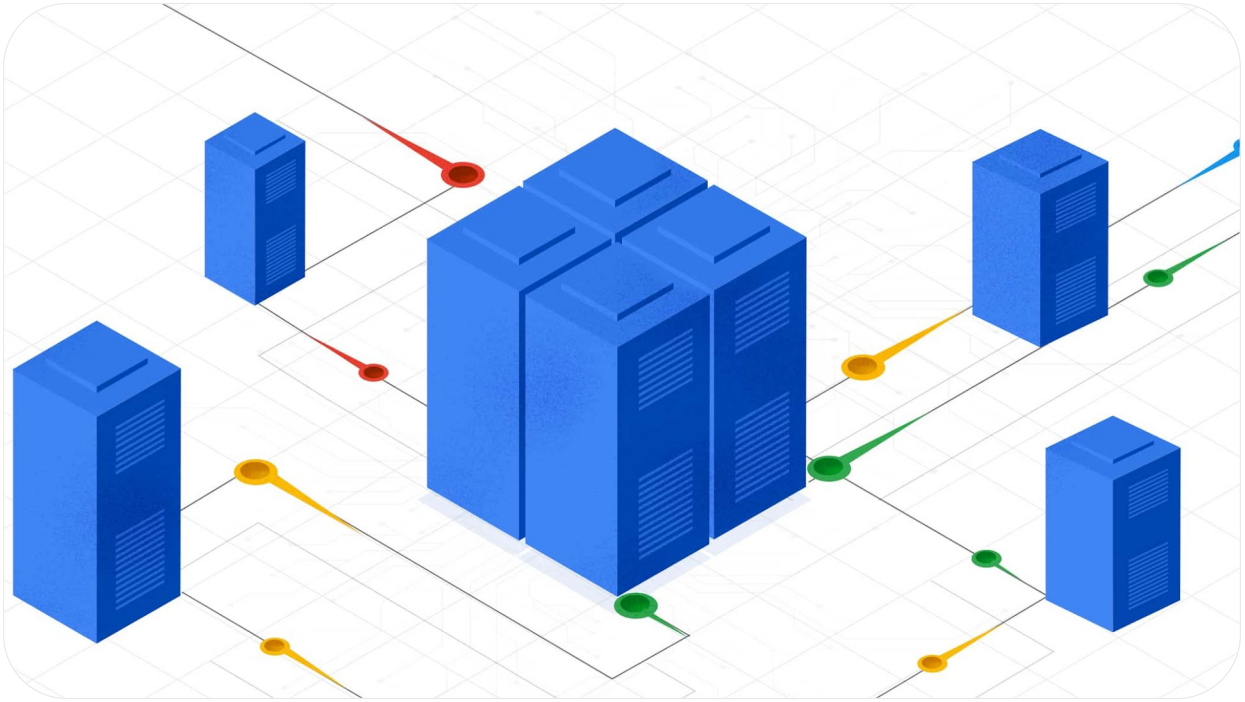
Storage & Data Transfer

Colossus under the hood: a peek into Google's scalable storage system

April 20, 2021



Dean Hildebrand
Technical Director, Office of the CTO, Google Cloud

Denis Serenyi
Tech Lead, Google Cloud Storage

same underlying storage infrastructure for its other businesses as well? That's right, the same storage system that powers Google Cloud also underpins Google's most popular products, supporting globally available services like YouTube, Drive, and Gmail.

That foundational storage system is Colossus, which backs Google's extensive ecosystem of storage services, such as Cloud Storage and Firestore, supporting a diverse range of workloads, including transaction processing, data serving, analytics and archiving, boot disks, and home directories.

In this post, we take a deeper look at the storage infrastructure behind your VMs, specifically the Colossus file system, and how it helps enable massive scalability and data durability for Google services as well as your applications.

Google Cloud scales because Google scales

storage services:

- **Colossus** is our cluster-level file system, successor to the Google File System (GFS).
- **Spanner** is our globally-consistent, scalable relational database.
- **Borg** is a scalable job scheduler that launches everything from compute to storage services. It was and continues to be a big influence on the design and development of Kubernetes.

These three core building blocks are used to provide the underlying infrastructure for all Google Cloud storage services, from Firestore to Cloud SQL to Filestore, and Cloud Storage. Whenever you access your favorite storage services, the same three building blocks are working together to provide everything you need. Borg provisions the needed resources, Spanner stores all the metadata about access permissions and data location, and then Colossus manages, stores, and provides access to all your data.

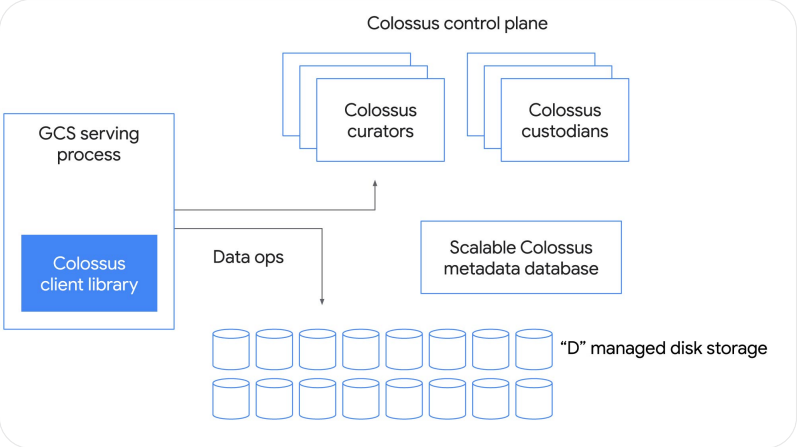
Google Cloud takes these same building blocks and then layers everything needed to provide the level of availability, performance, and durability you need from your storage services. In other words, your own applications will scale

Now, let's take a closer look at how Colossus works.

But first, a little background on Colossus:

- It's the next-generation of the GFS.
- Its design enhances storage scalability and improves availability to handle the massive growth in data needs of an ever-growing number of applications.
- Colossus introduced a distributed metadata model that delivered a more scalable and highly available metadata subsystem.

But how does it all work? And how can one file system underpin such a wide range of workloads? Below is a diagram of the key components of the Colossus control plane:



top of Colossus use a variety of encodings to fine-tune performance and cost trade-offs for different workloads.

Colossus Control Plane

The foundation of Colossus is its scalable metadata service, which consists of many Curators. Clients talk directly to curators for control operations, such as file creation, and can scale horizontally.

Metadata database

Curators store file system metadata in Google's high-performance NoSQL database, BigTable. The original motivation for building Colossus was to solve scaling limits we experienced with Google File System (GFS) when trying to accommodate metadata related to Search. Storing file metadata in BigTable allowed Colossus to scale up by over 100x over the largest GFS clusters.

D File Servers

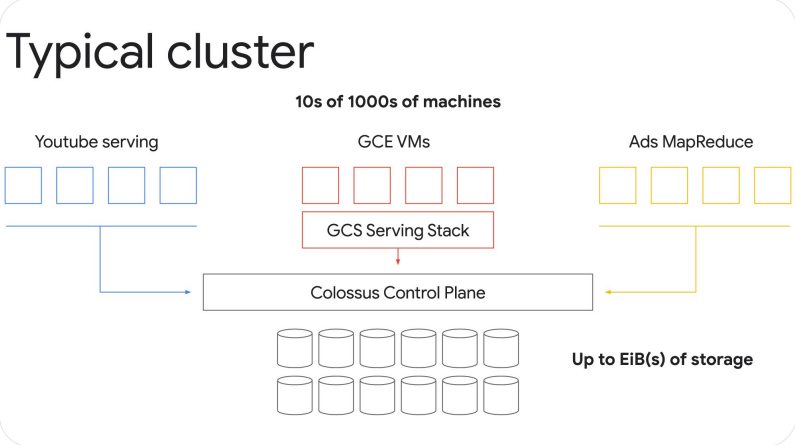
Colossus also minimizes the number of hops for data on the network. Data flows directly between clients and "D" file servers (our network attached disks).

Custodians

Colossus also includes background storage managers called Custodians. They play a key

How Colossus provides rock-solid, scalable storage

To see how this all works in action, let's consider how Cloud Storage uses Colossus. You've probably heard us talk a lot about how Cloud Storage can support a wide range of use cases, from archival storage to high throughput analytics, but we don't often talk about the system that lies beneath.



With Colossus, a single cluster is scalable to exabytes of storage and tens of thousands of machines. In the example above, for example, we have instances accessing Cloud Storage from Compute Engine VMs, YouTube serving nodes, and Ads MapReduce nodes—all of which

Disaggregation of resources drives more efficient use of valuable resources and lowers costs across all workloads. For instance, it's possible to provision for the peak demand of low latency workloads, like a YouTube video, and then run batch analytic workloads more cheaply by having them fill in the gaps of otherwise idle time.

Let's take a look at a few other benefits Colossus brings to the table.

Simplify hardware complexity

As you might imagine, any file system supporting Google services has fairly daunting throughput and scaling requirements that must handle multi-TB files and massive datasets. Colossus abstracts away a lot of physical hardware complexity that would otherwise plague storage-intensive applications.

Google data centers have a tremendous variety of underlying storage hardware, offering a mix of spinning disk and flash storage in many sizes and types. On top of this, applications have extremely diverse requirements around

units.

In addition, at Google scale, hardware is failing virtually all the time—not because it's unreliable, but because there's a lot of it. Failures are a natural part of operating at such an enormous scale, and it's imperative that its file system provide fault tolerance and transparent recovery. Colossus steers IO around these failures and does fast background recovery to provide highly durable and available storage.

The end result is that the associated complexity headaches of dealing with hardware resources are significantly reduced, making it easy for any application to get and use the storage it requires.

Maximize storage efficiency

Now, as you might imagine it takes some management magic to ensure that storage resources are available when applications need them without overprovisioning. Colossus takes advantage of the fact that data has a wide variety of access patterns and frequencies (i.e., hot data that is accessed frequently) and uses a

capacity. With the right mix, we can maximize storage efficiency and avoid wasteful overprovisioning.

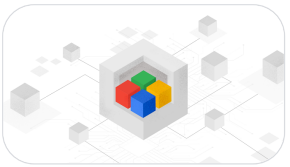
For disk-based storage, we want to keep disks full and busy to avoid excess inventory and wasted disk IOPs. To do this, Colossus uses intelligent disk management to get as much value as possible from available disk IOPs. Newly written data (i.e. hotter data) is evenly distributed across all the drives in a cluster. Data is then rebalanced and moved to larger capacity drives as it ages and becomes colder. This works great for analytics workloads, for example, where data typically cools off as it ages.

Battle-tested to deliver massive scale

So, there you have it—Colossus is the secret scaling superpower behind Google's storage infrastructure. Colossus not only handles the storage needs of Google Cloud services, but also provides the storage capabilities of Google's internal storage needs, helping to deliver content to the billions of people using Search, Maps, YouTube, and more every single

architecture, check out the Next '20 session from which this post was developed, "A peek at the Google Storage infrastructure behind the VM ." And check out the cloud storage website to learn more about all our storage offerings.

Storage & Data Transfer



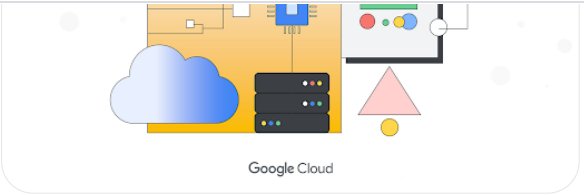
Optimizing object storage costs in Google Cloud: location and classes

Saving on Cloud Storage starts with picking the right storage for your use case, and making sure you follow best practices.

By Dom Zippilli • 7-minute read

Posted in Storage & Data Transfer—Google Cloud

Related articles



Infrastructure Modernization

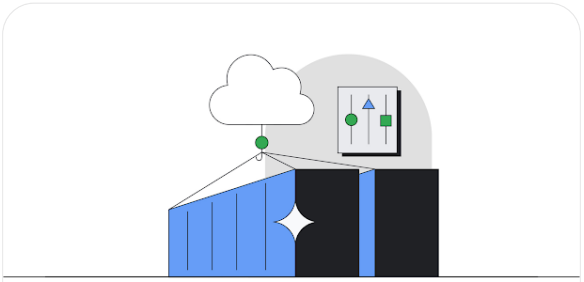
Google Cloud VMware Engine supercharged with Google Cloud NetApp Volumes

By Nirav Mehta • 4-minute read

Compute

Optimize costs for Windows workloads using Persistent Disk Async Replication

By Brian Kudzia • 4-minute read



Containers & Kubernetes

Leveraging Backup for GKE (BfG) for Effortless Volume Migration: From In-tree to CSI

By Arun Singh • 5-minute read

HPC

Salk Institute scientists scale brain research on Google Cloud with SkyPilot

By Qiurui Zeng • 7-minute read