Google Cloud

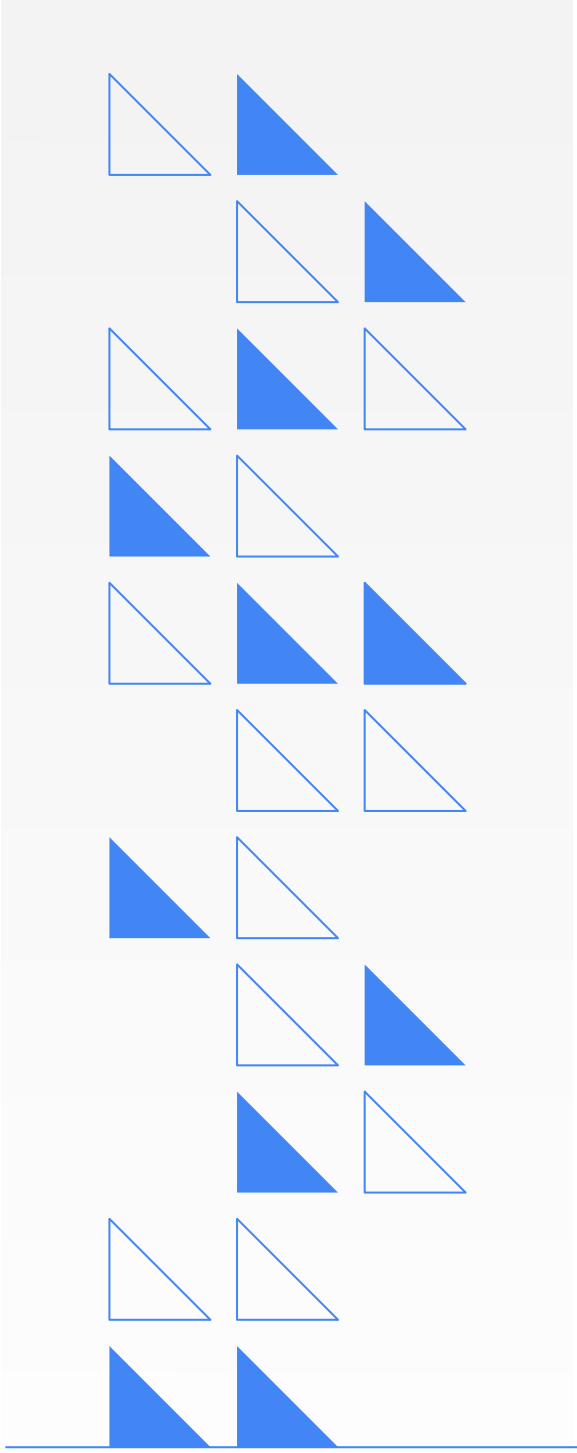
Maximize Cost Efficiency with CUD Optimization Strategy

August 2022

Authors

Sheri Cunningham, Eric Lam, Sarah McMullin, Rich Hoyer, Jason Gerhart, and Robin Roacho

Table of contents



By incorporating best practices and committing to use key sets of Google Cloud services, organizations can save up to 70% on certain cloud costs. To minimize costs, examine architectures and practices for opportunities to reduce Google Cloud consumption volumes. It's critical for cloud FinOps professionals and users of Google Cloud services to understand the distinction between operational efficiency and pricing efficiency.

Operational efficiency: managing volumes of consumption

Example: hours of runtime for compute resources

Pricing efficiency: negotiating private pricing for an organization

Example: committed use discount (CUD), enterprise discount

Ensuring operational efficiency before pricing efficiency helps avoid the temptation to reduce costs by participating in discount programs without first addressing the vast underutilization or overprovisioning of resources. This whitepaper introduces principles to improve operational efficiencies and examines strategies that drive cost savings using CUDs – a key component of maximizing pricing efficiency.

What are committed use discounts?

Google Cloud CUDs offer discounted prices in exchange for a commitment to use **either a minimum level of resources or to spend a minimum amount** for a specified duration term of one year or three years. Once committed, users are not able to cancel within their agreed upon term.

There are two types of CUDs:

- **Resource-based CUDs** provide a discount in exchange for a commitment to use a minimum level of Compute Engine resources in a particular region. The discount applies to vCPUs, memory, GPUs, and Local SSDs and typically varies from 37% to 57%. If users are actively using memory-optimized machine types, however, the discount can be as high as 70%. Users can purchase commitments on a **per-region basis** and not on a more restrictive zone level.

  CUDs do not apply to:

  - VMs created using the App Engine flexible environment or Dataflow
  - Spot VMs instances, shared-core N1 machine types, or extended memory
  - Projects that are within the free tier with **free tier** credits
  - Projects that do not have any payment history

Additionally, users can only buy a one-year commitment for K80 GPUs. CUD prices for committed resources are **independent of on-demand pricing**, and any fluctuation in on-demand pricing would not necessarily be reflected in CUDs.

Daily consumption levels below the commitment amount will result in some waste since the commitment forms the minimum cost per microsecond, also known as the commitment window, once a CUD is purchased.* **Unutilized commitments do not carry over from the commitment window**, making CUDs unsuitable for bursty workloads.

Example:

50 vCPUs for N2D machine series in region us-central1

Scoped to an individual project or Cloud Billing account if discount sharing is not enabled

- **Spend-based CUDs** provide a discount in exchange for a commitment to spend a minimum amount of money per hour for a product or service. If utilization is more than the commitment, the overage is charged at a regular on-demand rate. If utilization is less than the commitment, then optimal savings aren't being reached.

  Spend-based CUDs can be purchased to provide discounts for the following services:

  - Cloud SQL
  - Google Cloud VMware Engine
  - Cloud Run
  - Autopilot in Google Kubernetes Engine

Example:

$50/hour spend in Cloud SQL for PostgreSQL in us-central1

Applies to aggregated spend for resources that are scoped to all eligible projects paid by the billing account

* Overall cost may still be lower than if on-demand pricing were in place.

Comparison of CUD types

	Resource-based	Spend-based
What do users commit to?	Volume of consumption of applicable Compute Engine resources per microsecond	Hourly dollar spend in USD
Limited to a region?	Yes	Yes typically, but depends on the specific product
Available terms	1 year / 3 years	1 year / 3 years
Where is the CUD purchased?	In the Compute Engine console for a cloud project	On the Cloud Billing page of a Cloud Billing account
Scoped to	Cloud project, or Cloud Billing account if committed use discount sharing is enabled	Cloud Billing account
Services	Compute Engine, GKE, Dataproc (vCPUs, memory, GPUs, some OS licenses, and Local SSDs)	Cloud SQL, VMware Engine, Cloud Run, GKE (Autopilot mode), Cloud Spanner
Offers auto-renew option?	Yes	No

Why purchase CUDs?

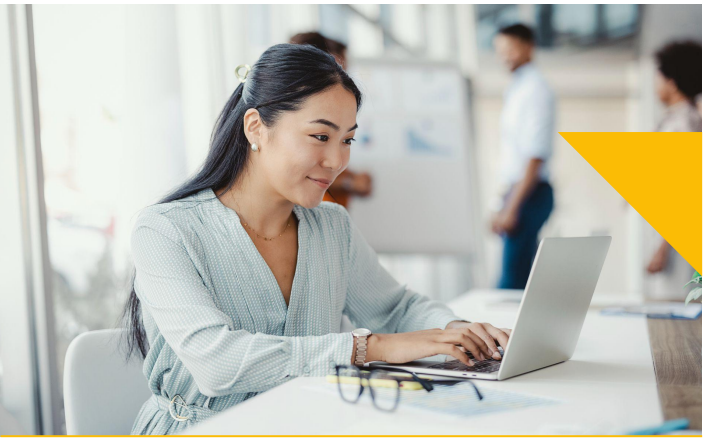
Ideally, organizations will implement a cost-savings strategy that incorporates the optimization of both operational and pricing efficiencies. CUDs are the easiest of the discount mechanisms for users to take advantage of, since they can be purchased with a single click. The organization's CUD coverage also can be adjusted on an ongoing basis as consumption patterns change.

Splitting and merging CUDs

To help users manage resource requirements for projects, Compute Engine offers the ability to merge or split existing commitments and redistribute resources to match the granularity required for projects.

Splitting CUDs

Users can transfer resources out of an existing commitment, also known as a source commitment, and split the commitment into two smaller commitments, or split commitments. Splitting allows users to closely monitor and manage portions of one large commitment in the form of smaller individual commitments. For example, users can set only a portion of a commitment to renew automatically by splitting it and enabling automatic renewal for only one of the split commitments. Users can also use the splitting functionality in conjunction with prioritized attribution of commitments to decide how to distribute CUDs at a more granular level.



Requirements for splitting

When users split a source commitment to create two split commitments, they must meet the following requirements:

- The new split commitment must have the same project, commitment type, region, and duration term as the source commitment, but users can choose a new name for the split commitment.
- The resource types specified for the new split commitment must match some or all of the resource types in the source commitment. Additionally, the amount of resources specified for the new split commitment must be a portion of the resources in the source commitment. For example, if the original source commitment is for 200 vCPUs and 300 GB memory, then the split commitment can have 0–200 vCPUs and 0–300 GB memory, but cannot have both 200 vCPUs and 300 GB memory.
- The source and split commitments must be for hardware resources like vCPUs, memory, GPUs, and Local SSDs.

Merging CUDs

Users can merge multiple active CUDs to create a new larger commitment. This will result in a new commitment with the end date corresponding to the latest end date from the merged commitments and will include the sum of the original resources in a single commitment. For example, if a user merges commit A, which expires on January 2, 2023, with commit B, which expires on April 10, 2023, then the merged commitment will expire on April 10, 2023.

By merging commitments together, users can track and manage them as a single entity. Merging commitments helps users avoid staggered commitment expiration dates by getting the individual commitments to expire at the same time. Merging also allows users to gradually ramp up workloads. For example, users can purchase multiple smaller commitments when the need arises and then either merge them together or with an existing commitment.

Requirements for merging

When users merge individual source commitments to create a newly merged commitment, the source and merged commitments must meet the following requirements:

- All of the source commitments must have the same project, region, duration term, commitment type, and commitment category.
- The newly merged commitment must have the same project, region, duration term, commitment type, and commitment category as the source commitments, but users can choose a new name for the merged commitment.
- The resource types users specify for a merged commitment must be the exact same that are in the source commitments. Additionally, the amount of resources for each resource type in the newly merged commitment must be equal to the sum of the amounts of resources for that resource type in all the source commitments. For example, if the first source commitment has 100 vCPUs and 100 GB memory and the second source commitment has 200 vCPUs and 300 GB memory, then the user must create a merged commitment with 300 vCPUs and 400 GB memory.
- The source and merged commitments must be for hardware resources like vCPUs, memory, GPUs, and Local SSDs.

By default, when users create merged or split commitments, the auto-renew setting is disabled on the new commitments – even if all the source commitments were set to renew automatically. If users want merged or split commitments to renew automatically, they must manually enable the auto-renew setting on those commitments. Users can do this either at the time of creation or after.

If a user performs a merge or split of CUDs and has prioritized attribution enabled, then all existing attributions for those commitments will be deleted. The user will have to reconfigure prioritized attribution on the resulting commitments.

Limitations to splitting and merging commitments are as follows:

- Users can’t merge or split commitments using the Google Cloud console UI
- Users can’t merge or split license commitments
- Users can’t split a commitment when it has an attached reservation
- Users can’t split commitments for GPUs or Local SSDs, since these resources must have reservations attached to them at all times
- At the time of creation of merged or split commitments, users can’t create any new reservations and attach them to those commitments
- Users can’t merge or split commitments that have expired or have been canceled

Best strategies to leverage CUDs

CUDs are ideal for workloads with predictable and stable usage patterns. There's no option to choose both spend-based and resource-based CUDs for the same resource, so there's no associated optimization strategy.

Choosing the right CUD duration term: one year or three years

CUDs are offered in either one- or three-year durations and the discount is higher for three-year CUDs. For each contemplated CUD resource purchase, practitioners should confirm the duration of their planned usage and decide if they're likely to make changes in their usage patterns that would make a one-year commitment preferable. This is particularly true for resource-based CUDs, which are specific to both a product and a region.

When usage of a product is highly concentrated among a small number of technical owners, they should be directly consulted by the FinOps practitioners managing the CUD portfolio about their upcoming plans. If the usage is highly dispersed among many technical owners, then leaders in the IT organization can help provide insights on the time frame of planned changes in usage patterns.

CUD sharing across projects

Google Cloud users have the ability to turn on CUD sharing, which allows the discount to be shared across all projects tied to a billing account. This discount is proportional by default. So if a project uses 60% of its total resources, it will get 60% of the CUD discount.

As a general practice, it's recommended to share CUDs among the broadest possible range of accounts and projects, since sharing begets the highest savings for an organization. To concentrate their savings, some technical owners may advocate for the ability to buy CUDs that don't use sharing and thus are limited to specific accounts or projects. If the utilization in specific projects is projected to be stable or to sustain long-term growth while that of the rest of the organization is highly volatile, then users should still enable CUD sharing while utilizing prioritized attribution to allocate savings. Attribution for CUDs determines how fees and credits are applied to a Cloud Billing account and spread across the account's eligible projects. This is reflected in how the fees and credits appear on the Cloud Billing Reports page, accessed through the usage cost export or the Google Cloud console. In this example, the stable projects can still achieve their full savings and any remaining savings won't be wasted, but instead applied to lower-prioritized projects.

Checking for concentrated usage

Users should identify how concentrated the usage of a product is within specific teams at the organization before buying associated CUDs. Low concentrations of use among technical owners make it less likely that the organization will make changes in consumption that are drastic enough to preclude a CUD purchase – while products with higher concentrations of use will require more caution. If the concentration is low, the cloud FinOps user doesn't need to approach individual users of the service, but instead should approach the cloud engineering or cloud infrastructure director or cloud infrastructure general manager to verify that no dramatic changes in consumption of products or regions are contemplated during the duration of the proposed CUD purchase.

If the concentration is high, where over 70% of the usage is attributed to one to three teams, FinOps Practitioners should reach out to the parties generating the spend and gather information on their usage plans for the duration of the contemplated CUD purchase. Key questions to ask are:

- Is the current level of spending expected to be higher, remain constant, or decline over the next one to three years?
- Would the technical owner see any reason not to purchase a commitment based on any other factors, such as the potential for moving workloads to different regions?

The answers to these questions will guide cloud FinOps practitioners on how much usage should be covered with CUDs and what CUD durations make the most sense for the instance types.

Resource-based CUDs

Prepurchase items to check for

We recommend limiting CUD coverage on specific Google Cloud products or services to a modest level until confirming the following:

- The projected consumption of the products is expected to remain at or above current levels for the duration of the CUD. Always-on usage, rather than volatile, is more predictable.
- For the products for which CUDs are contemplated, typical rightsizing recommendations do not result in significant cost reduction.
- The usage of each CUD SKU is not highly concentrated within a small number of technical owners. Although high concentration will not preclude the CUD purchase, plans for continued use need to be confirmed beforehand when usage is concentrated.

Best practices for leveraging resource-based CUDs

Once the CUD analysis is complete, users can begin making purchases that will gradually work toward a target CUD coverage rate of 80%. Making purchases in stages allows the user to monitor utilization and ensure that there are no unexpected operating conditions like highly volatile consumption that may result in CUD waste. If such patterns are encountered, additional purchases can be suspended to maintain a low enough CUD coverage rate.

If either of the last two conditions from the analysis is false, purchasing CUDs to target 25% initial coverage of the current spend would be a safe purchase until rightsizing recommendations and concentrated usage are checked.

Checking for rightsizing recommendations

The Google Cloud console provides recommendations to change compute instance types or delete them altogether when they're underutilized. The recommendations are based on the trailing eight-day period of CPU or memory usage and can be found in the Compute Engine recommendations tab of the console. If recommendations to downsize instances exist for contemplated CUD SKUs, a review of rightsizing recommendations is warranted before significant commitments are made, since resources could be changed to other SKUs or terminated altogether as part of the rightsizing process. Once rightsizing recommendations have been executed or ruled out, the purchase recommendations can be re-evaluated against the revised spend and purchases can be made.

Spend-based CUDs

Prepurchase items to check for

The process for refining purchases of spend-based CUDs is similar to that of resource-based CUDs with one exception: Checking for rightsizing recommendations is less relevant, since spend-based CUDs do not cover specific SKUs, but rather many individual SKUs within a service offering. Ideally, users should limit their CUD coverage on specific products to modest levels until after they have confirmed the following conditions:

- The projected consumption of the products is expected to remain at or above current levels for the duration of the CUD. Always-on usage, rather than volatile, is preferred.
- The usage of each CUD SKU is not highly concentrated within a small number of technical owners.

If both conditions are true, users can begin making purchases that will gradually work toward a target CUD coverage rate of 80%. Making purchases in stages allows the user to monitor utilization and ensure that there are no unforeseen operating conditions, such as highly volatile consumption that may result in CUD waste. If such patterns are encountered, additional purchases can be suspended to maintain a low enough CUD coverage rate to avoid waste.

If the second condition is false, purchasing CUDs to target 25% initial coverage of the current spend can still be a safe purchase while discussing usage with the concentration of users.

Customer success story

An internet branding company was spending about $175,000 per month on compute RAM and CPU across N1, E2, and N2 instance types. Although they were able to reduce their costs to some degree by using Spot VM instances, most of their workloads weren't tolerant of being preempted on short notice and needed to run on demand. They were eager to find ways to reduce their expenses on compute instances outside of increasing their Spot VM consumption. CUDs were the natural choice.

Their cloud FinOps practitioners generated a spend report that segmented their spend on compute by the organization's technical owners who were responsible for operating them. Their analysis showed that while the usage of N1s and E2s were spread fairly evenly across a number of technical owners, usage of N2s was highly concentrated within two teams. Before making the CUD purchases, the practitioners approached the heavy users of N2s and discovered the users planned to maintain or grow their use of N2s over the next several years.

Next, FinOps practitioners accessed the Recommender API to determine whether a large enough number of recommendations were in place to cause reductions or changes in their usage patterns. After overseeing implementation of some of the recommendations, they generated a forecast of how much the remaining recommendations might reduce their E2 and N1 consumption in the future, adjusting their spend forecasts accordingly.

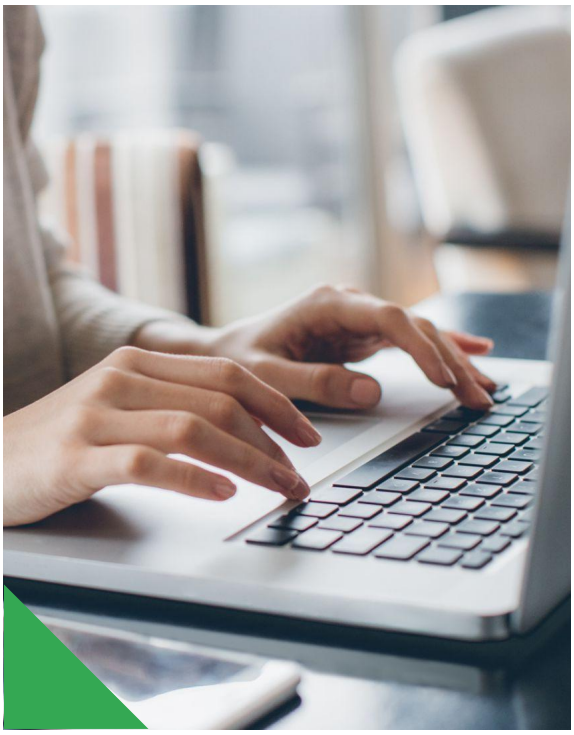
Feeling confident that they now had a conservative and reliable forecast for consumption, the cloud FinOps team submitted a request for their management team to approve purchase of the CUDs. After being granted approval, they purchased enough three-year CUDs to cover roughly 95% of their forecasted consumption. As a result, the internet branding company reduced their costs on compute RAM and CPU by over 50%.

How to purchase and analyze CUDs

To further explore cost savings **via resource-based CUDs** with Google Cloud, see Purchasing commitments for machine types or Purchasing commitments for GPUs and local SSDs. If a reservation is attached to a CUD, the reservation cannot be deleted for the duration of the commitment.

After purchase, users are billed a monthly fee for the duration of the term selected – whether or not the services are used. Take a walkthrough of utilizing Cloud Billing reports to Analyze the effectiveness of resource-based CUDs.

To learn more about and purchase **spend-based CUDs** for Cloud SQL, VMware Engine, Cloud Run, or Autopilot mode in GKE, see Purchasing spend-based commitments. For reporting and analysis of current commitments in place, see Analyze the effectiveness of spend-based commitments.



CUD recommender

The CUD recommender helps users optimize the resource costs of the projects in their Cloud Billing account. Its recommendations are generated automatically based on historical usage metrics gathered by Cloud Billing. Users can use these recommendations to purchase additional commitments and further optimize their Google Cloud costs.

Recommendations are available for both spend-based and resource-based commitments for Cloud Billing accounts billed in US dollars.

The CUD recommender provides two types of recommendations:

- **Stable usage recommendation is** the more conservative option and determines the commitment amount of CUDs that will cover the minimum stable usage for the most recent 30 days.
- **Optimal savings recommendation determines** the optimal commitment amount for the most recent 30 days of overall usage, also known as the look-back window, where usage covered by the recommended commitment is more cost-effective than usage at list or on-demand prices. Note that this may mean that the commitment recommendation modeled on the look-back window might cover resources that are not active all of the time, but the net effect is an overall cost savings based on the overall amount of uptime.

Monitoring utilization

After CUD purchases have been made, CUD utilization should be monitored for two weeks to verify that usage patterns are not generating waste or vacancy. If the initial purchases were conservatively made and resulted in relatively low CUD coverage, additional purchases should be made to work up to the target of 80% coverage for stable or growing workloads. Each time a purchase is contemplated, technical owners for highly concentrated workloads and IT leaders for dispersed workloads should be consulted about the purchase to verify that plans have not changed.

Organizations needing additional help with their CUD strategy can reach out to either their Google Cloud account team or SADA for professional services. Google Cloud or SADA will generate procedures customized to the organization's specific combination of Google Cloud services and cost-management tooling.

Special thanks to Yasmin Mowafy and Charan Kuchibhotla for providing their domain expertise and continuous support on this important cloud FinOps topic.