



FIVE POWERFUL COST OPTIMISATION LEVERS IN GOOGLE CLOUD



Welcome to “Five Powerful Cost Optimisation Levers in Google Cloud”

In 2023, organisations spent roughly \$564 billion on the cloud. When surveyed, most organisations estimated that around 30% of that cloud spend was wasted, meaning a whopping \$169 billion of waste — the most in any year, ever.

In 2024, organisations are set to spend nearly \$700 billion on the cloud.

Are you getting the maximum value from your cloud spend?

If you're looking for a way to immediately reduce wasted spend and carbon emissions in Google Cloud then look no further. Here are five key levels you can pull right away to impress your peers, your bosses, and your friends ;)



Hi, I'm **Andrew Tate**, one of co-founders of **StackPartners**. We're a group of **Google Cloud Engineering experts** with a passion for delivering value for customers. For over 15 years I've been building and running business-critical platforms and products and ensuring maximum return on technology investments.

But why are we telling you this? Because we understand how hard this stuff is, we've been in your shoes. Whether it's the difficulty in creating an accurate cloud cost forecast, the inability to find time for an effective cost optimisation initiative, lack of visibility and accountability for cloud spend, the uncertainty of how much cloud spend you're wasting, challenges in creating a cost-conscious culture across technology and product teams, overspending without awareness, slow to address cost anomalies, lack of deep experience in Google Cloud, teams too busy firefighting or not engaged in cost optimisation or sustainability engineering, the complexity of legacy applications and databases - we've lived it all. That is why we started [StackPartners](#), to make your and your team's lives easier and help you thrive in Google Cloud for years to come.

“Five Powerful Cost Optimisation Levers in Google Cloud” contains some of our key learnings that you and your teams can use right away to optimise your Google Cloud costs.

Rightsizing your Google Kubernetes Engine (GKE) Clusters

There are some great ways to save a lot of money (and the environment) when it comes to rightsizing. One of our favourites for Google Kubernetes Engine (GKE) is to look at the max pods per node.

Rightsizing workloads in GKE can be challenging, here's a few things you can check and implement to ensure you're not wasting a load of resources and cash. Ensure you match the instance size/type in the worker node pool to the max-pods-per-node setting. If you have n2-standard-8 type worker nodes but a very low max pods per node setting e.g. 16 then the clusters will scale up not because of the need for more CPU and/or memory, but because new pods can't be scheduled on existing nodes that already have 16 pods running. This results in lots of CPU and memory resources being provisioned but not being utilised. Increasing the max-pods-per-node to 32 means it's possible to schedule more pods per node, utilising all available CPU and memory. This results in fewer nodes being required for the same workload. With this one tweak, we saved a customer around \$70K a year!

Understanding the resource requirements of workloads in GKE is essential for cost-effective instance selection. Each workload has specific needs for CPU, memory, storage, and network throughput. You need to accurately assess these needs, so you can choose the most appropriate instance types and prevent over or under-provisioning, which both lead to wasted costs and performance issues. Properly matching instance types to workload requirements ensures efficient resource usage, optimising for both performance and cost. This careful alignment is key to maintaining cost-effectiveness and scalability, especially in a microservices architecture.

Logging Optimisation

Reducing or adjusting the sampling rate of logs is an effective strategy for saving on cloud logging costs. By selectively capturing a representative subset of log data rather than logging every single event, organisations can significantly decrease the volume of stored data, thereby cutting down on storage and retrieval expenses. This approach ensures that critical insights are still obtained while avoiding the costs associated with excessive log data. Implementing dynamic sampling rates, where logging frequency is adjusted based on the significance or frequency of events, further optimises costs without compromising the quality of monitoring and debugging capabilities.

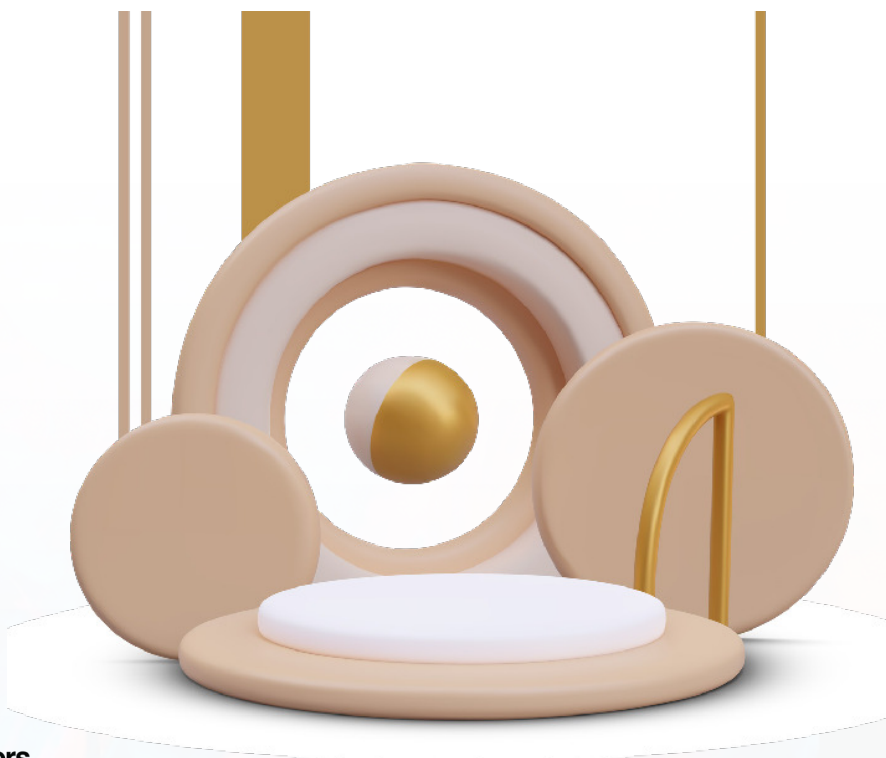
In a real-world a GKE network logging with a 100% sampling rate may generate approximately 1.88TB of VPC flow logs. However, by reducing the sampling rate to 25%, the amount of log data stored can be significantly decreased to “only” 480GB. This reduction illustrates how adjusting the sampling rate can lead to substantial savings on cloud logging costs while maintaining adequate visibility into network activities.

VPC Flow Logs 100% Sampling Rate = 1.88TB a month = \$11940 a year

VPC Flow Logs 25% Sampling Rate = 480GB a month = \$2748 a year

That is a saving of \$9k a year and that’s only a single subnet.

Logging management can be a juggling act to balance observability, costs, and security. Achieving this balance requires careful consideration of the volume and detail of logs collected to ensure sufficient insights without incurring excessive expenses. A transparent dialogue with all stakeholders—product teams, DevOps, and security—is paramount to managing this effectively. By collaborating and understanding each group’s needs and priorities, organisations can develop a logging strategy that meets operational requirements, optimises costs, and upholds security standards.



3












Committed Use Discounts

Committed use discounts (CUDs) provide savings in exchange for your commitment to use a minimum level of resources for a specified term, either one or three years. These discounts should be used for workloads with predictable traffic patterns/resource requirements. You will get recommendations on which CUDs to purchase directly in the Google Cloud billing console, do it now! Well actually, do it after you've read the below.

In Google Cloud, there are two types of CUDs; spend-based and resource-based.

Spend-based committed use discounts provide a discount in exchange for your commitment to spend a minimum amount over the term. You must purchase separate spend-based commitments for each Google Cloud service that you want discounts for. The committed use discounts apply to usage of the services in any project and any region paid for by the Cloud Billing account.

You can get spend-based committed use discounts for the following Google Cloud services:

-  **AlloyDB for PostgreSQL**
-  **Backup and DR Service**
-  **Bigtable**
-  **Cloud Run**
-  **Dataflow**
-  **Spanner**
-  **Cloud SQL**
-  **Compute Engine (Can be used for Google Kubernetes Engine)**
-  **Google Cloud VMware Engine**
-  **Google Kubernetes Engine (Autopilot)**
-  **Memorystore**

Resource-based committed use discounts provide a discount in exchange for your commitment to use a minimum level of a specific Compute Engine resource in a particular region. The discount can be purchased for vCPUs, memory, GPUs, and local SSDs. You must specify the resource type you are committing to e.g. General-purpose N2, europe-west2. You can enable discount sharing so that the Compute Engine committed use discounts are shared across all projects that are paid for by the same Cloud Billing account.



4

Negotiated Savings

If you're planning to ramp up spending over a couple of years and/or can commit to spending mid to high 6-figure numbers per year, you're able to negotiate a discount off the on-demand list price of services. This will come in the form of a percentage discount and more often than not a large amount of credits that are transferred to your Google Cloud billing account. Your other discounts i.e. CUDs and sustained use discounts will be applied on top of these negotiated savings. Most (but not all) services count towards the committed spend and usually, at most half of the committed spend amount can come from Google Marketplace spend.



5

Use the FinOps Hub Recommendations

In the Google Cloud console go to your billing page, click on the FinOps Hub under Cost Optimisation, under Potential savings/month click View all recommendations. Check the tabs "Switch off idle resources" "Right size instances" and "Purchase CUDs." Do what it says on the tin and you'll be winning in no time.

If you're looking for a way to reduce wasted spend and carbon emissions so you can thrive in Google Cloud for years to come. [Sign up to our wait list here and we'll be in touch.](#)

Solid foundations make a strong house

Effective cloud cost management begins with a solid organisational structure that lays the foundation for financial accountability and operational efficiency. This structure involves clearly defined roles and responsibilities, ensuring that everyone from engineers to finance teams understands their part in managing cloud expenses.

By establishing governance frameworks and policies, organisations can set guidelines for resource usage, budgeting, and cost allocation, promoting a culture of cost-consciousness. Proper tagging and labelling of resources enable accurate tracking and reporting of expenses, facilitating informed decision-making and accountability across departments.


Furthermore, integrating FinOps practices into the organisational workflow encourages collaboration between technical and financial teams, fostering an environment where cloud spending is continuously monitored, optimised, and aligned with business objectives. This structured approach not only helps in controlling costs but also maximises the value derived from cloud investments, driving overall business success.

In a multi-tenant platform leveraging Google Kubernetes Engine (GKE), effective cloud cost management can be achieved through the strategic use of namespaces, tagging, and labelling. Here's an example to illustrate this approach: **Example:** Multi-Tenant SaaS Platform on GKE


Organisational Structure


A software-as-a-service (SaaS) company provides a platform used by multiple teams. Each team's workload is isolated within its namespace in a shared GKE cluster. The company aims to manage cloud costs effectively by tracking and optimizing resource usage for each team.


Use of GKE Namespaces


 **Namespaces for Isolation:** Each team has a dedicated namespace (e.g., teamA-namespace, teamB-namespace). This setup ensures logical separation and easier management of resources and policies.


Tagging and Labeling Strategy


 **Resource Labels:** All Kubernetes resources (e.g., pods, services, deployments) within each namespace are labeled with relevant information. Common labels include:


 **app:** <application-name> (e.g., app: billing-service)


 **env:** <environment> (e.g., env: production, env: staging)

 **team:** <team-name> (e.g., team: teamA)

 **Namespace Annotations:** Each namespace is annotated with metadata for cost tracking and organizational context. Examples include:

 **team:** <team-name> (e.g., team: sales)

 **department:** <department-name> (e.g., department: marketing)

 **cost-center:** <cost-center-id> (e.g., cost-center: CC1234)

Billing Export and GKE Cost Allocation

Understanding cloud costs is critical for effective financial management in Google Cloud. When setting up Google Cloud, it is essential to create the billing export immediately, as it only exports data from the day of its setup. By establishing this export early, you ensure comprehensive tracking and analysis of your cloud expenditures from the outset. Utilizing this billing export, you can visualise and monitor your Google Cloud costs, enabling you to identify spending patterns, optimise resource allocation, and make informed decisions to manage and control your cloud budget effectively.

GKE cost allocation is a powerful tool for managing and optimising infrastructure costs in a multi-tenanted environment. By providing detailed insights into resource usage at the namespace level, it enables precise cost allocation and accountability among different teams or projects. Development teams can use this data to understand their consumption patterns and take ownership of their infrastructure costs, promoting a culture of cost-awareness and responsibility. This transparency empowers teams to make informed decisions about resource optimisation, scaling, and budgeting, ultimately leading to more efficient and cost-effective cloud infrastructure management.

Reinvestment Ideas

You have some options for using the money you saved, simply return it to your bottom line or...

Research and Development (R&D): Allocate funds to explore new technologies, develop new products, or improve existing ones.

Prototyping and Experimentation: Invest in experimental projects or prototypes that can lead to breakthrough innovations.

Feature Development: Accelerate the development of new features or enhancements to your existing products, increasing their value to customers.

Performance Optimization: Invest in optimising the performance and scalability of your products.

Green Projects: Invest in environmentally friendly projects or technologies to reduce your company's carbon footprint.

Energy Efficiency: Improve the energy efficiency of your operations, potentially reducing future costs.

Security Tools and Services: Invest in advanced security tools and services to enhance your cloud security posture.

Compliance Audits: Ensure compliance with industry standards and regulations through regular audits and certifications.

Tag [StackPartners](#) or give us a mention if this helped you and let us know how much you saved!

If you're looking for a way to reduce wasted spend and carbon emissions so you can thrive in Google Cloud for years to come. [Sign up to our wait list here and we'll be in touch.](#)