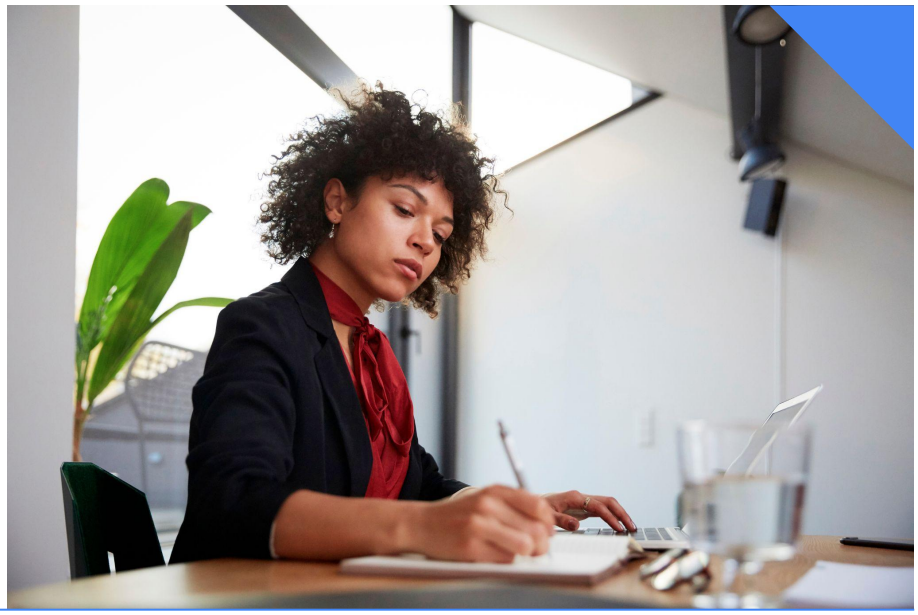
Google Cloud

Cloud FinOps - Shared Services Cost Allocation

October 2022



Kundan Sarkar - Customer Engineer

Srinivas Cheemalapati - Cloud Customer Experience Lead

Alberto Paez - Big Data and Analytics Cloud Consultant

Special thanks to Bruce Warner, Eric Lam, and Pathik Sharma for contributing and providing their domain expertise and continuous support to this important cloud FinOps topic.

Introduction

As organizations adopt cloud services to run their operations efficiently, it's imperative for business leaders to understand the true cost of such services and identify sources of cost overrun. One cost element that most enterprises seek to understand is the cost of shared cloud services. When the multitude of cloud services offered by the cloud providers are considered, it is very likely that some cloud services costs are shared across multiple lines of business (LoB) and often the challenge is how to allocate such costs across these groups. Developing a precise means to allocate shared cloud costs improves the transparency and the efficiency of cloud cost management across IT, Finance, and lines of business. It also increases the accountability of the business users and acts as a catalyst to better align with the IT organization.

This white paper discusses the foundational elements of cost allocation and the complexities and challenges associated with shared services cost allocation. While some of these concepts and models are applicable interchangeably between legacy and cloud environments, this white paper focuses primarily on cloud computing and its associated services.

Understanding Showback vs Chargeback

The two most common cost allocation strategies used by IT organizations to improve alignment and accountability are Showback and Chargeback. Showback is a process by which an IT organization provides visibility into IT resource usage without charging its customers (LoBs) for their use. Chargeback is a process by which an IT organization can recover its cloud cost of providing services to its customers (LoBs).

Implementing showback and chargeback models are often iterative in nature and almost always a part of the organization's cloud finops crawl-walk-run journey. The table below summarizes recommended models for each phase and typical activities being performed by the model:

Maturity phase	Crawl	Walk	Run
Model	Visibility	Showback	Chargeback
Activities	• Cloud usage reporting • Tracking of cloud spend by products / services • Monthly trend analysis and forecasting • Cloud billing details by LoBs • Application Total Cost of Ownership • Shared services cost allocation • Establish data structure (account, folder and tagging design)	• Cloud Unit costing • Cloud services definition • Cost allocation mapping to business services • Cost metrics reporting • Tagging aligned to business P&L • Dashboard / reporting by LoBs	• LoBs accountable for cloud services cost • Chargeback based on actual consumption • Cloud services catalog for service provisioning • Automation of cloud usage monitoring

Adopting Showback and Chargeback models benefits organization operations in many different ways as outlined below:

Cost transparency : Increased transparency due to accurate allocation of cloud costs and usage

Awareness : Increased awareness of how IT contributes to the business of the organization

Alignment : Cloud costs become integrated into business planning, strategy, and operations

Accountability : Responsibility of cloud costs shifts to the LoBs

Cost efficiency : Provides the basis to develop unit costs, which in turn drive better architectural and consumption decisions

Financial planning : Improved demand and financial forecasting by enabling business leaders a clearer image of cloud costs

It is important to note here that Chargeback sends expenses to a product or department P&L and Showback shows the charges by product or department but keeps the expenses in a centralized budget. So it is imperative that once the chargeback has been implemented, it needs to be integrated with the organization's internal reporting systems and financial management tools.

Direct Costs vs Shared Services Costs

Both Showback and Chargeback models are based on two primary cost components - *Direct costs* and *Shared Services costs*.

Direct costs are those that can be attributed specifically to a desired measurement group. This could be a line of business, a product or service, a specific end-user customer, or an application as examples. Regardless of the measurement group, there is a 1-1 relationship between the provisioned cloud resource, such as a compute instance, database, or storage bucket, and the discrete measurement group

Shared service costs are those that cannot be attributed specifically to a desired measurement group, where specific cloud services are utilized across multiple cost objects. Examples include multiple lines of business consuming a common kubernetes resource, a database instance used by multiple products or services, or a storage resource leveraged across multiple end-user customers. Allocating these costs fairly and objectively across all business users is often a challenge.

The essential difference between direct costs and shared services costs is that only direct costs can be traced to specific cost objects.

Shared Costs Personas and Challenges

A Cloud Finops practice brings together stakeholders from the organization's business, technology and finance groups and they often have conflicting goals and objectives to meet. While lines of business strive to grow the top line revenue, finance typically assesses financial performance and drives capital allocation and investment decisions and engineering focuses on meeting SLAs.

Some examples of the personas and their priorities are outlined below:

Business

Program Owner / Platform Manager

- Business value
- Regional compliance
- Faster time to market
- Competitive advantages
- Growth strategies

Finance

Controller

- Budget and Forecast accuracy
- Cost granularity
- Cost structure and competitiveness
- Consumption accountability with chargeback and allocation
- Cost amortization
- Profitability

Technology

Engineering Manager / Developer

- Mapping LoBs and applications to projects
- Tagging and labeling of resources
- Toolings / recommendations for cost optimization
- Scalability of infrastructure to meet growth projections

Shared cost allocation technique should consider this complexity of different personas and their specific challenges while defining and implementing the appropriate model for the organization.

Shared Costs Types

Now, let's look at different types of shared costs that IT organizations are commonly dealing with:

- Enterprise discounts
- Enterprise licenses
- Organizational functions (Cloud Center of Excellence, Site Reliability Engineering, PMO office, etc.)
- Technology (Data lake, Data pipeline, Innovation with AI/ML later consumed by multiple teams, etc)
- Shared resources (Network, Storage, Databases, Cloud Operations, Compute)
- Support
- Marketplace services

In a typical cloud deployment, all of the above cost categories (and there could be many others based on how cloud infrastructure is provisioned versus how cost objects are defined and measured) could be shared across multiple LoBs, applications and environments. For an example scenario, consider a compute cloud service, where different LoBs can qualify for additional discounts when there is aggregation of the consumption. In this scenario, the shared cost allocation model should be able to derive the consumption ratios of different LoBs and apply consumption discounts to each of these LoBs proportionately.

Steps to Implement Shared Cost Allocation

The key to any successful cost allocation is to establish an allocation system that is fair, equitable, and supported by current data. At high level, steps to implement shared cost allocation models are :

Step 1 :
Identify the cloud services costs to be allocated

Step 2 :
Categorize costs by direct costs vs. shared services costs

Step 3 :
Determine the allocation approach and method to distribute each type of shared costs equitably

Step 4 :
Allocate the shared services costs by one of the cost allocation methods

Step 5 :
Update and monitor the data and methodology to ensure the allocation remains fair and equitable over time

Cloud Cost allocation methods are very dynamic in nature. As the organization is constantly improving its understanding and usage of the cloud, the service provider is also constantly improving the product and adding more cost optimization and cost visibility features into the services. These directly impact the correctness and accuracy of the allocation method and the constant updating and monitoring of the data and the methodology is the best practice recommendation.

Shared Cost Allocation Models

Many costs that go into the total cost of ownership(TCO) calculation are shared, making it difficult to assign specific cloud infrastructure costs to a single application, LoBs, service, or capability. To handle these costs, businesses need a fair cost allocation method, similar to activity based costing used for manufacturing where activities (metrics) are defined that drive IT spending and use these metrics as a basis for assigning costs within a tier, or in moving costs from one tier to another.

Typically shared cost allocation methods are broadly categorized into Attribution based allocation method and Consumption based allocation method.

Attribution based methods can be further subdivided into:

Proportional

Shared costs are allocated based on the direct cost proportion

Even Split

Shared costs are allocated evenly across LoBs

Fixed

Shared costs are allocated based on business-defined coefficient (the sum of coefficients needs to be 100%)

Consumption based methods systematically identify cloud services consumption at the individual component level to drive cost allocation e.g using API calls to measure relative use of an application.

Implementation of an attribution-based model

Most **Shared Services Costs** are addressed by **attribution**. The reference cloud services implementation model for this example scenario is shown in figure 1. It is organized in a typical hierarchical manner with resources consumed by projects at the bottom, platforms are built on top of the projects, applications then consume platform resources and finally business users access applications. There could be projects shared between platforms as well. Support costs and alike will come under shared costs.

As mentioned above, the attribution distribution can be classified into three different categories:

Proportional: In this method, shared services costs will be distributed among the LoBs based on the proportion of the direct costs incurred. In this model all the costs are rolled up through the hierarchy to the LoBs by adding direct costs and proportional costs for shared projects and/or platforms. In cases where the direct costs are shared between LoBs they are apportioned based on some percentage that the organization decides.

Even Split: Under the even split model, the shared services (ex: support costs) will be shared evenly by all LoBs. Because of its simplicity, this model tends to be more popular in smaller organizations that have fewer LoBs. In the reference model the total rollup of the organizational spend is evenly divided between the LoBs. A much simpler example will be to have 5 tenants using a GCP resource and assign 20% of the cost for each.

Fixed: In this model, shared costs are allocated based on business-defined coefficient (the sum of coefficients needs to be 100%). The fixed proportion method relies on using a set percentage to attribute the direct and shared costs for the overall spend month over month. Typically these ratios have been determined by evaluating past spend and arriving at a fair breakdown for allocating spend. For example, a percentage of total services used by each LoB or function.

These models can be explained quantitatively using a simple example shown in Table 1, with 3 LoB projects whose Direct costs are identified. The networking costs (ex: Ingress, Egress, VPC, DNS, CDN etc) categorized as shared costs are obtained. The total direct costs across the three projects is $100K and the total shared networking cost is $30K.

In the proportional model using total direct cost as the denominator, B1-P1 is allocated 50% of shared networking costs ($15K) and similarly B2-P2 gets 30% of shared networking costs ($9K) and so on. The total cost is just the addition of the associated project direct cost and the allocated cost.

In the Even split model, the shared networking cost of $30K is evenly split across the three LoB projects ($10K) and then added to the Direct costs per project to get the total cost.

In the Fixed split model, a predetermined ratio is applied at the project level for the shared costs. As the example shows, the weightage of 3:2:1 is applied to the shared networking costs ($15K, $10K, $5K respectively) and the totals are calculated by projects. The key takeaway is that the total cost varies by the model and depends on how the organization chooses to implement.

LOB - Projects	B1-P1	B2-P2	B3-P3
Direct costs	$50K	$30K	$20K
Indirect cost - Networking		$30K	

	Proportional split (based on direct costs)		Even split		Fixed split (based on predefined ratio of 3:2:1)	
LOB-projects	Allocated cost	Total cost	Allocated cost	Total cost	Allocated cost	Total cost
B1-P1	$15K	$65K	$10K	$60K	$15K	$65K
B2-P2	$9K	$39K	$10K	$40K	$10K	$40K
B3-P3	$6K	$26K	$10K	$30K	$5K	$25K

Table 1: Example of Shared Cost Allocation across LoB projects

Best practices indicate that the more granular that labels and resources are assigned, the easier it is to associate cloud costs to specific cost objects.

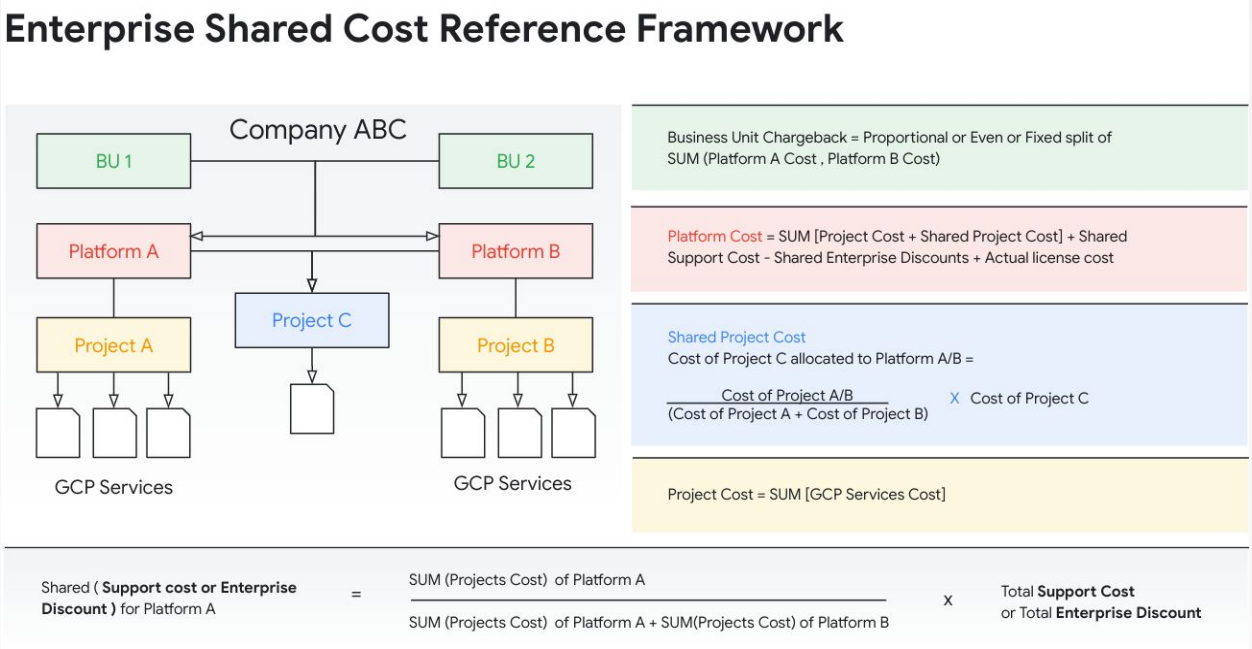


Fig 1: Shared Cost Reference Framework

The above diagram represents a reference framework to allocate shared costs of cloud consumption to business units. Let's discuss the framework in the context of an online retailer. Different business units (BU1 and BU2) are owning the retailer's custom applications and packaged applications. To support these applications, the IT team grouped their functions logically under Platforms where Platform A represents the Web platform and Platform B represents the Data platform. While applications are using both platforms, certain GCP services are specific to platforms, e.g. Platform A uses GKE whereas Platform B uses Cloud Spanner. These specific services are grouped within specific projects, Project A and Project B, for Platform A and B respectively. There are additional GCP services that are common to both platforms, e.g. Load Balancing, Networking, etc and these services are grouped together in Project C.

Shared costs of Project C are allocated across Platforms A and B using one of the attribution based models (Proportional, Even or Fixed) adopted by the organization. Similarly, other shared costs across platforms , e.g. support costs, enterprise discounts, license costs etc are also allocated based on the attribution model of choice by the organization. Finally the showback/chargeback to respective business units are calculated adding these platform costs and applying the attribution-based model (proportion,even or fixed) used by the organization.

Implementation of a consumption-based model

Consumption-based models systematically identify cloud services consumption at the individual component level to drive cost allocation. Cloud labeling is the technique used to capture the consumption. A label is a customer-defined key and value pair used to organize Google Cloud resources. Labeling allows the organization to enable categorization of resources by purpose, owner, environment,etc.

In Google Cloud, labels are attached to resources and all resources are organized in a resource hierarchy. The purpose of the hierarchy is two-fold:

- Provide a hierarchy of ownership
- Provide inheritance for access control and organization policies

The diagram to the right is an example of a typical resource hierarchy in an organization.

The Organization resource is at the top of the hierarchy and all resources that belong to an organization are grouped under the organization resource.Folder resources are optional logical resources to model different legal entities, departments and teams across the organization.

Project resources are base level organization entities. It forms the basis of creating, enabling and using all Google Cloud services, managing APIs, enabling billing and managing permissions.Cloud resources like Compute, Storage, Databases are created within projects and it is with these resources that the labels are attached.

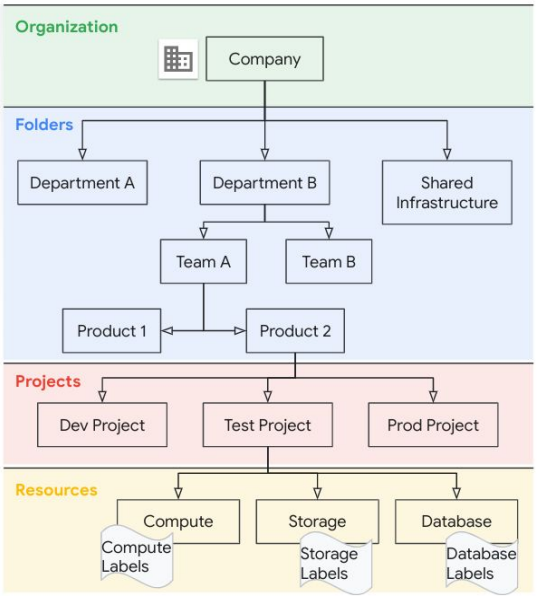


Fig 2: Sample Resource Hierarchy

Cloud labeling framework defines business objectives and metrics across four key dimensions:

- **Technical labels** - owner, enddate, environment, function, release, etc
- **Business labels** - application, cost center, project id, business units, departments, etc
- **Security labels** - compliance, encryption, classification, etc
- **Automation labels** - enddate, schedule-off, schedule-on, criticality, etc

Key steps to implement cloud labels are as follows:

- Ask critical questions
- Develop labeling strategy
- Assess in-scope resources
- Enable showback/chargeback model
- Adjust and optimize

For a much granular level of tracking cloud usage at the user and the transaction level where labeling is not available, it is recommended to extract the information from the Cloud Operations Suite logging data or other metadata dictionaries (such as INFORMATION_SCHEMA for BigQuery)

Conclusion

In this white paper, we have discussed some frameworks to allocate shared costs of cloud consumption across users. These frameworks are generic in nature and need to be curated from one organization to the next based on its operating policies and organization structure. One should also keep in mind that there will be considerable differences in approach and timeline when it comes to adoption of these frameworks by a legacy organization vis a vis a cloud native organization. Also, the concepts and models discussed here , particularly the Consumption based model, are based on the current functionalities of how the billing captures services consumptions and may change in future with additions of new features and functions. And lastly, it's important to consider that cost management is a continuous effort and process, defined strategies and approaches should be reviewed to ensure it's still aligned with business functioning and costing strategy.