Deloitte.

Frugal Cloud:
How to better manage cloud costs

February 2024

Introduction

> JR Storment, co-author of the book Cloud FinOps, Collaborative, Real-Time Cloud Financial Management, once said, The dirty little secret of cloud spending is that the bill never goes down.[1]

It seems cloud service and implementation providers worldwide took up a challenge to prove the quote wrong, and thus the concept of Cloud Financial Management (CFM) was born.

CFM, as the term suggests, is about embracing the scale, agility, and innovation provided while reigning in the costs associated with the cloud.[2]

The concept has gained immense popularity now, primarily due to the number of sleepless nights CIOs globally are having to ensure cloud costs are within manageable limits.

On one hand, a recent global **Deloitte survey**[3] suggests that 90 percent of surveyed cloud decision-makers believe that cloud servers are the force multiplier for their digital strategy. On the other hand, another **survey conducted by Tangoe**[4] (a US-based cloud expense management solution firm) indicates that 80 percent of surveyed decision-makers consider cloud cost savings as their biggest issue in achieving benefits from existing cloud deployments.

Let us dig a little deeper into why cloud decision-makers are finding it difficult to realise cloud savings.

1. Lack of understanding of cloud economics

Organisations and relevant cloud stakeholders have little or no idea about the economics related to cloud with some underlying challenges:

- Reading and understanding **complex cloud contracts** is a gargantuan task for many infrastructure and operations leads. Cloud providers use billing models and pricing structures with myriad options and combinations that lead to a lack of understanding in selecting the best pricing option.
- **Complex cloud discounting structures** and payment schedules lead to confusion amongst consumers and under-utilisation of discounting mechanisms.
- **Basic cost hygiene protocols** are seldom placed as foundational cornerstones for building the cloud marvel.

Ways to improve cloud economics understanding

It is important for companies to hire advisors/subject matter experts (on contract/payroll) for:

Pre- and post-contract reviews to evaluate potential accounting treatment impacts and develop audit-ready accounting analysis

Ascertaining various discounts that can be used to reduce procurement costs. For example, commitment-based discounts, sustained use discounts, universal credits, discounts available on usage of pre-emptiable/reserved/spot machines, and bring-your-own licence

Helping compare various discounts across dimensions such as ease of purchase, ease of changes, ability to cancel, payment options, and other options to combine with

Helping procure the best-fit cloud financial management tools to increase spend transparency, optimise the cloud spend, manage financial operations, and govern cloud sprawl

2. Unforeseen/Unpredicted usage

Unprecedented focus on customer centricity and the ever-growing demand for customer insights have put the power in the hands of developers, data scientists, and data engineers who are tapping into the massive compute and storage capability of the cloud to churn out these AI/ML models, data products (part of data mesh), and customer-facing apps and platforms to enable organisations to gallop ahead of competition . In many such scenarios, an oversight of cloud usage is sacrificed at the altar of competitive advantage, thereby blowing out original budget estimates due to untapped, unmeasured cloud usage. Tech projects built on cloud resources often go unnoticed and unmeasured on parameters of benefit realisation, eventually leading to organisational angst at the end of the year. It is imperative for firms to maintain a close watch on cloud spend transparency, which encapsulates two focal dimensions:

- **Visibility:** Cost of ownership across business units and show-back/chargeback
  - According to the 2021 annual State of FinOps report[5] survey from the FinOps Foundation, dealing with shared cloud costs was rated as the second-biggest challenge for firms trying to reduce their cloud spend, with a whopping 33 percent of respondents claiming they do not split shared cloud costs. Whereas in 2022, accurate predictions of cloud spend were the second-biggest, CFM-related challenge. The two challenges are tied to the same umbilical cord of a root cause, which is the lack of cloud spend visibility across various lines of business in a firm. This lack of clear cost apportionment and accountability often fuels the unexpected usage of cloud resources, manifesting in the form of the proliferation of cloud apps, platforms, and data products.
- **Performance management:** Granular drilldowns into cloud spending
  - Predictable cloud usage entails a granular know-how of cloud burning, and slicing and dicing of consumption-related data into multidimensional and panoramic views. This further requires efficient state-of-the-art cloud financial management tools. The State of FinOps survey5 suggests that one-third of respondents continue to use native tooling provided by cloud service providers, which, though serving a lot of reporting and data visualisation needs, can at times prove to be extremely foundational.
  - In many cases, organisations manually track cloud spend and implement chargeback/show-back mechanisms, a fact bolstered by the State of FinOps Foundation survey,[5] wherein 33 percent of respondents claimed to implement these mechanisms using spreadsheets.

How to manage unforeseen/unpredicted usage

Visibility

- Two important cost-attribution models can be deployed:
  - Assumption-based model, i.e., equal distribution of costs/allocation based on proportion of total spend/based on the number of resources supporting each unit
  - Consumption-based model, i.e., amount of core cloud services consumed by each unit/number of resources supporting each unit
- It is also important to consolidate indirect and direct costs, book them into financial systems, and charge various business units, business P&Ls, and cost centres for cloud usage.

Performance Mismanagement

- Strategic FinOps management platforms/solutions can be deployed, which can provide granular visibility and insights into resource utilisation by various lines of business in a multi-cloud setup. These solutions come with flexible cloud cost reporting options to meet the varying needs of decision-makers, thereby helping firms implement and automate chargeback/show-back mechanisms.
- The tooling pyramid structure can be implemented with the use of:
  - Foundational CSP-provided tools by application owners/engineers/architects
  - Third-party cloud financial management tools by functional/domain managers/reporting managers
  - High-power, scalable data visualisation tools by executive-level leaders for fine-grained reporting and drilldowns

3. Lack/Mismanagement of commitment

Unpredictable cloud usage and a lack of clear usage patterns have, over time, made the idea of commitment-based negotiated discounts a pipe dream for many firms. Firms do not prefer commitment-based contracting with cloud service providers as decision-makers are unsure about the use of cloud services in the future. Organisations with predictable demand and traffic spikes must negotiate commitment-based discounts with CSPs.

Firms that purchase commitments often do little to manage the committed portfolio, as explained in the following points:

- There is little focus on changing **reserved instance configurations** as the cloud footprint evolves.
- Identifying **areas of overcommitment** and changing modifiable commitment attributes during low-utilisation periods becomes a second priority.
- Firms invest little in **commitment automation tools** that provide the flexibility to combine multiple commitment models to drive maximum savings.

How to improve commitment

Committing too much or too little can lead an organisation to overspend. In some cases, it is prudent to overcommit to drive down overall costs compared with the pay-as-you-go mode. Meanwhile, in others, overcommitment can drive complacency in the firm and drive capacity optimisation initiatives. Various strategies can be put into place to improve the management of commitment portfolio:

Commitment automation, i.e., process of managing and optimising cloud resources and commitments through automated tools and processes

Licensing optimisation, i.e., use of special licensing programmes offered by cloud service providers that enable clients to bring their licences to the cloud without having to pay again

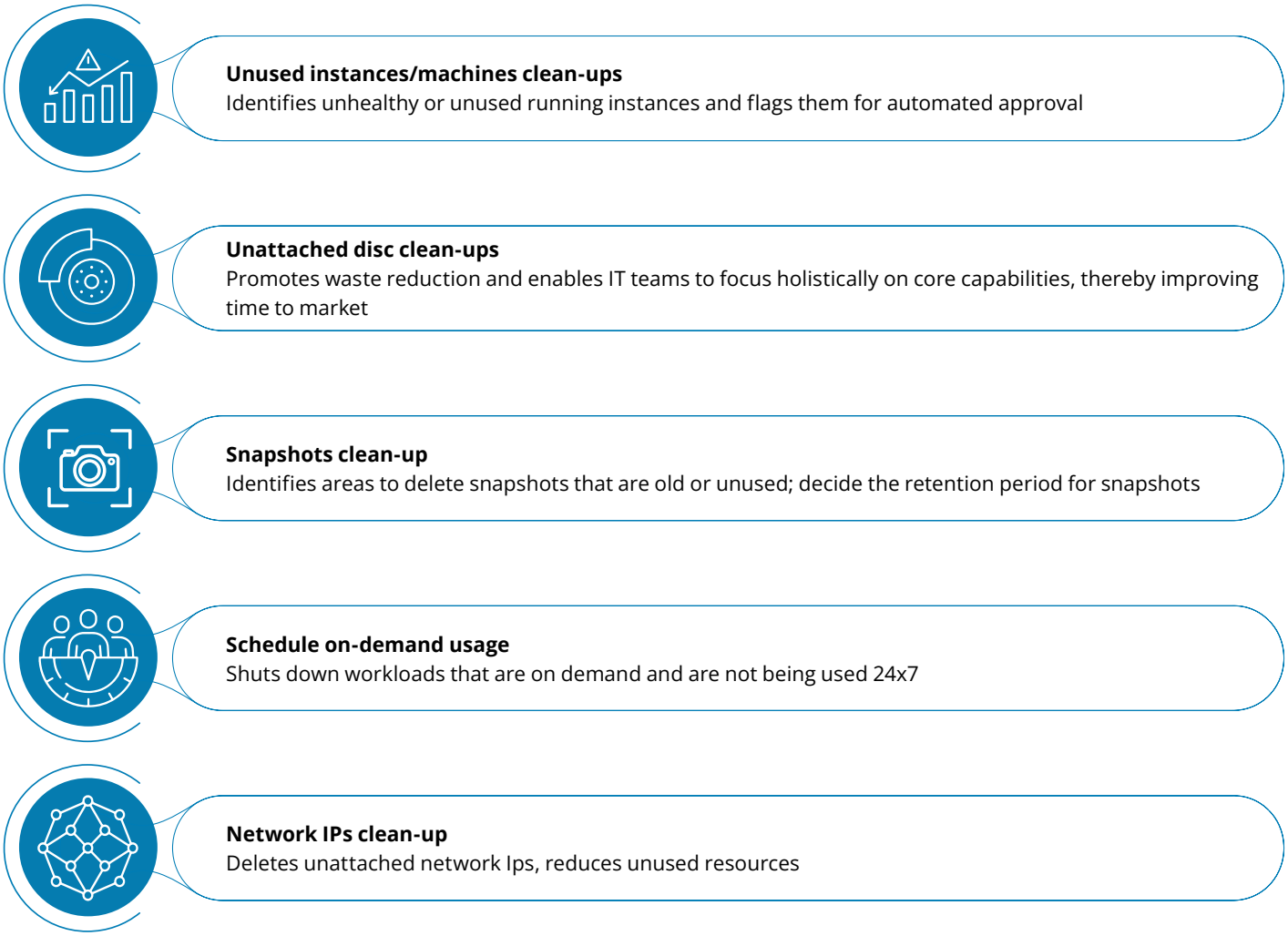
Alteration in the use of instances based on applicable footprint and utilisation

Firms also need to use various categories of cloud discounts to achieve maximum possible gains.

#	Category	Discount description
1	Ease of purchase	Factors influencing the purchase of a commitment-based instance: • Reserved instances/machines based on term • Region instance/machine type • # of vCPUs • # of GB RAM • Network type
2	Payment options	All up front, partial up front, no up front, monthly payments, pay as you go
3	Ability to cancel	Ability to cancel a committed purchase of an instance; credits could either expire, a penalty could be charged, or they could be exchanged in the marketplace
4	Ease of change	The ability to change a committed purchase; the change could be either for a different instance type or a different region

4. Consumption wastage

According to Deloitte's cloud financial management experts' experience, the top five ways to reduce cloud consumption waste are mentioned below:



Unused instances/machines clean-ups
Identifies unhealthy or unused running instances and flags them for automated approval

Unattached disc clean-ups
Promotes waste reduction and enables IT teams to focus holistically on core capabilities, thereby improving time to market

Snapshots clean-up
Identifies areas to delete snapshots that are old or unused; decide the retention period for snapshots

Schedule on-demand usage
Shuts down workloads that are on demand and are not being used 24x7

Network IPs clean-up
Deletes unattached network Ips, reduces unused resources

Another major contributor to cloud waste is the inability to right-size cloud resources; in fact, the FinOps Foundation survey[5] suggests right-sizing of instances as the third most important priority amongst the survey respondents. Firms have made successful journeys to the cloud, but the on-prem resource deployment and usage mindset will take some more years to change. Production environments are overprovisioned and forgotten in traditional data centres due to a comfortable capex spend, which is planned for at the start of every year. Provisioning of resources works differently on the cloud, where firms pay for the provisioned resources monthly. Thus, it comes down to rightsizing every machine on the cloud as per the desired requirement.

Right-sizing machines is tricky given the plethora of options and associated constraints provided by CSPs. In the case of complex cloud services with complex billing models such as container services, data and analytics platforms, serverless services, and managed services, the problems get aggravated. To resolve rightsizing-related complexities, it is important for firms to:

- Use best-fit cloud financial management tools with features related to downsizing/optimisation recommendations
- Baseline instance performance before setting the rightsizing initiative into motion, and then monitor and measure performance after rightsizing execution
- Bring back the sizing to the original baseline in case of a performance downgrade (cloud is a saviour when it comes to up-scaling/down-scaling)

For complex services, a mix and match of strategies can be deployed

- **Data and analytics services:** Configuration changes and changes in data modelling can prove helpful in optimising utilisation.
- **Container services:** Use of a master-slave arrangement of nodes for running containers can help upscale and downscale containers, as in this arrangement, addition and elimination of nodes from clusters are easy.
- **Serverless services:** Minimising external calls, eliminating/reducing job-polling or task coordination, and benchmarking serverless services for under/overutilisation of what the function is allocated can be deployed as an optimisation technique

Ways to reduce waste

Firms can use various levers to reduce cloud waste. Some of these levers include the following:

Instance rightsizing

Analyse current instances and determine the correct instance size for optimum cost and performance

Automating optimisation

Using automation to create alerts and report on waste and trigger automated actions to resolve inefficiencies

Tagging strategy

Apply organisation-wide tagging policies to enable better tracking and visibility into account-level spending patterns

Expensive storage replacement

Select the correct cloud storage and prevent overspending

Data transfer cost optimisation

Identify savings opportunities by placing workloads where data transfer costs can be optimised

Case study

A large US healthcare firm had a significant footprint on the cloud and ongoing migration to enable a multi-cloud strategy. Hence, it needed to mature a cloud financial management practice to manage and optimise cloud spend effectively.

Deloitte provided robust cloud financial management capabilities through FinOps-as-a-Service catalogue. Within the first month, it could identify 30 percent monthly savings; quick wins saved US$15 million in annualised cloud spend.

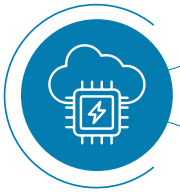
5. Legacy application architectures

Modernisation/refactoring applications in the cloud can help achieve equal or better cost savings than those delivered by waste and consumption levers. These cost savings come concomitant with other benefits of modernisation, such as application agility and application development acceleration. Firms can consider the following architecture modernisation strategies to achieve cost efficiencies:

Use of serverless/managed services:
Cloud's serverless services come with more fine-grained billing compared with server-based services, thereby providing better transparency and predictability on usage. Applications that can fit into the constraints of serverless services, such as stateless design, payload size limits, invocation time restrictions, and cold start challenges (code download followed by new environment execution for running serverless services), can be considered suitable candidates for re-architecting serverless patterns.

The use of platform-as-a-service components in the form of managed databases and cloud-native load balancers, especially for applications that have been rehosted on the cloud, can help reduce cloud bills.

Auto scaling: Architecting applications to be stateless, distributed, loosely coupled, and with load-balancing capabilities are some prerequisites for auto-scaling to work in the cloud. Auto-scaling helps eliminate over-provisioning of cloud resources, thereby reducing cloud costs.



Microservices architecture: Reusing code in the form of microservices to execute various functionalities of an application can further release cost efficiencies.

Network optimisation: Reducing active-active configurations, caching data, reducing use of public IPs, limiting deployments to single or maximum two cost-effective regions, and employing content delivery network services are some of the best practices for setting up a lean and cost-effective cloud network architecture.

Instance and data storage optimisation: Use of pre-emptible instances, along with object storage (for storing older data), while designing lean cloud-native application architectures.

Ways to modernise legacy assets

Modernisation of legacy assets to make them modern cloud native and thereby reduce costs is a process comprising steps executed in a piecemeal manner. Various components of the application are modernised in stages, with each module executed independently or in conjunction with other modules.

An example of a modernisation journey can be:

#	Component	Stage 1	Stage 2	Stage 3	Stage 4
1	User interface	Basic browser UI: HTML 5 Redesign	Adaptive: Specific designs for supported devices	Responsive: complex, single design	**Progressive webapps**
2	Business logic	**Components:** Refactor based on OO principles: Separation of concerns, interfaces, rules engine	**Services:** Expose programme functions as a service	**APIs:** Expose business logic, rules, and workflow	**Microservices**
3	Framework	JPA Access Layer	Remove legacy emulation	**Modern frameworks** (Micronaut/Spring)	
4	Batch	Replacement of JCL	Batch to nearline	**Event-based processing**	
5	Data	Cloud-based data stores	Operational data stores	Strategic data stores: Data lakes	**Machine learning and AI**
6	Infrastructure	On-prem virtualisation	Containers	**Cloud-native/ serverless**	

Case study

A leading private insurance firm in India had rehosted its entire application estate on the cloud; it was struggling to reduce spiraling cloud costs.

Deloitte used its comprehensive cloud cost financial management offering to recommend large cost optimisation opportunities related to instance and storage resizing, and re-architecting applications to be more cloud native (i.e., use of load balancers, managed services, and serverless services).

Deloitte identified an annual cost savings of US$ 1 million as part of the CFM strategy

6. Absence of cloud saving culture

Fostering a culture of cloud cost savings in the organisation can prove to be an uphill task for the leadership due to the following reasons:

Lack of involvement of business: Business leads consider the cloud to be a strategic game changer for them but still believe it to be an IT initiative. Only when cloud costs become prohibitive to sustain will business leaders from finance, procurement, and strategy get involved, searching for faults and scapegoats. Cloud comes with its own sharp learning curve, and the late involvement of business makes the learning curve steeper. It is critical for firms to make business leaders cloud literate and have them involved in cloud matters and governance from the start.

Lack of CFM skills and capability: According to the FinOps Foundation 2023 survey, 47 percent of respondents believe their firm invests in cloud financial management training and education[8]. CFM skills and capabilities, such as those related to cloud usage and economics, cloud resource optimisation, and predictive analytics, are the foundational pillars for delivering enterprise-wide cloud savings.

Focus on tactical changes: The survey also suggests tactical CFM initiatives related to tagging, labelling, and hierarchy management are still priority areas of focus, with 9 out of 10 respondents claiming it to be a focus on CFM capability, whereas strategic initiatives related to unit economics and finance integration still continue to be the laggards, clearly pointing towards the misplaced priorities of firms when it comes to strategies for managing cloud costs.

Lack of clear metrics: Lack of structured, achievable success metrics related to cloud is a common challenge across multiple organisations, hindering the path to achieving cost efficiencies. It is important to implement a cloud economics-related KPI hierarchy within the firm, ranging from foundational KPIs such as cloud burn rate to complex KPIs such as segment prediction accuracy percentage.

Lack of motivation to act: The FinOps survey[5] suggests that motivating engineers to act on cost optimisation-related measures remained the top challenge over the past years, including this year. About ~30 percent respondents classified it as one of the top challenges. A few strategies can help induce this motivation in cloud engineers:

- **Gamification:** Internal competitions amongst engineering teams with representation on leaderboards for teams achieving the highest targets of cloud reduction
- **Incentivisation:** Tangible rewards for teams ending up as top performers on the leaderboards
- **SRE analogy:** Analogous to the Site Reliability Engineering (SRE) practice, in cases of deterioration of application performance, the engineering team shifts focus from new development to performance tuning and deployment. If cloud costs related to an application breach a certain threshold, the team puts the development of new features on hold and prioritises changes in application architecture and deployment to bring cloud costs under the specified threshold.

Ways to nurture cloud saving culture

Establishment of a Cloud Center of Excellence (CCoE) is one of the first foundational steps an organisation can take to nurture a cloud-saving culture. The CCoE can play a pivotal role in enterprise cloud competency assessments, including CFM-related skills assessments to help define workforce composition strategy, roles, and skills sourcing strategy (build/borrow/buy), capability development roadmaps, transition, and change management programmes. These initiatives can prove to be crucial in fostering a culture where cloud saving becomes a day-to-day priority.

Case study

A leading South Africa-based telco had cloud-first ambition but lacked relevant cloud skills and capabilities to manage cloud resources and keep cloud costs low.

As part of the CCoE offering, Deloitte India conducted several deep-dive sessions to assess CFM-related skills and competencies and define a comprehensive CFM capability development roadmap. The result was a cloud-first vision and strategy and a FinOps approach tailored to the client's specific needs.

These six key reasons (amongst various other reasons) have so far posed serious challenges to firms in reducing their cost [you mean cloud?] related costs; however, light is not far away at the end of the tunnel. Firms supported by the entire ecosystem of cloud service providers, cloud system integrators, and regulatory authorities, are embarking on the cloud financial management journey to mitigate these challenges. Though many of them are on the initial steps of this journey, taking these initial steps is a milestone in itself.

> Martin Luther King Jr. said, "You don't have to see the whole staircase, just take the first step."[6]

End notes

1. https://www.datacenterfrontier.com/featured/article/11428025/finops-the-growing-need-for-cloud-cost-crunching
2. https://docs.aws.amazon.com/wellarchitected/latest/management-and-governance-guide/cloudfinancialmanagement.html
3. https://www.tangoe.com/blog/cio-com-research-says-companies-face-a-cloud-cost-quandary/
4. https://www2.deloitte.com/us/en/pages/consulting/articles/cloud-insights-reporting-dashboard.html
5. https://data.finops.org/ | https://www.finops.org/introduction/how-to-use/
6. https://www.goodreads.com/quotes/1063592-you-don-t-have-to-see-the-whole-staircase-just-take

Contributor

Nidheesh Hirwani

Deloitte.