CAPCO

UNCOVERING THE OPTIMIZATION POTENTIAL OF CLOUD

STOP WASTING A THIRD OF YOUR SPEND

a **wipro** company

INTRO

Cloud adoption is today a priority for most enterprises[1] – but how best to optimize deployment of the cloud is too often overlooked in the rush to migrate. In particular, insufficient focus is placed on accelerating levels of cloud waste resulting from unused or under-utilized cloud resources or services. Without adequate forward planning, research and governance, cloud can become a more costly proposition than necessary, mitigating the gains such a transition promises.

Research on cloud migration[2] highlights cloud cost optimization as the leading cloud initiative for 2022 across organizations, yet organizations are estimated to waste one-third[3] of their cloud spend. So how is cloud waste generated, and what can organisations do to minimise it?

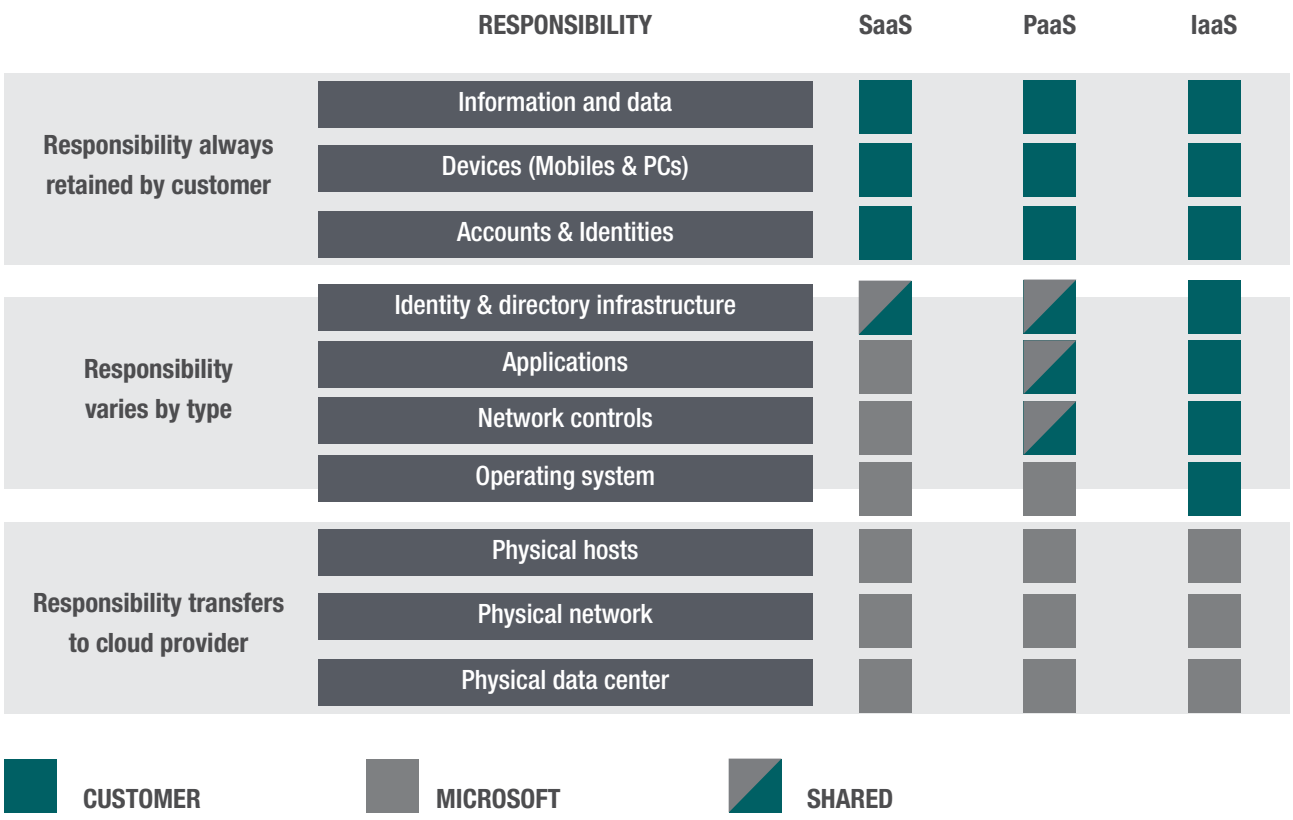
In its recent forecast of cloud spending[4] , Gartner identified the following areas:

	2021	2022	2023
Cloud Business Process Services (BPaaS)	51,410	55,598	60,619
Cloud Application Infrastructure Services (PaaS)	86,943	109,623	136,404
Cloud Application Services (SaaS)	152,184	176,622	208,080
Cloud Management and Security Services	26,665	30,471	35,218
Cloud Systems Infrastructure Services (IaaS)	91,642	119,717	156,276
Desktop as a Service (DaaS)	2,072	2,623	3,244
Total Market	**410,915**	**494,654**	**599,840**

Worldwide Public Cloud Services End-User Spending Forecast (Millions of U.S. Dollars)

Whilst waste can happen in any area of cloud spending, Infrastructure as a Service (IaaS) workloads – forecast to represent US$119.7 billion of expenditure in 2022 – which can potentially be an area of major waste.

Every cloud provider publishes its shared responsibility model across their offerings (Infrastructure, Platforms, Services etc). The highest responsibility for managing cloud resources lies with the customer (cloud consumers) when using Infrastructure as a Service (IaaS). This gradually decreases when moving towards Software as a Service (SaaS). Following is a representation of this from Microsoft Azure[5]. More the responsibility lies with the cloud consumer, higher is the risk of cloud waste – the more elements a customer has to manage, the higher the 'risk' of unoptimised usage, resulting in a higher risk of waste.



SOURCES OF CLOUD WASTE ([6,7])

The four main causes of waste are:

- Idle resources
- Overprovisioned resources
- Poorly managed storage
- Improper network topology

IDLE RESOURCES

Any resource in cloud when not in use is considered idle. A significant – in many cases the largest – element of compute spend is dedicated to non-production infrastructure or activities, which are typically used for only 40 hours a week. If those resources are shut down for the remaining unused 128 hours each week, significant cost savings would be made. However, shutting down an entire infrastructure and then bringing it back up when needed can be a huge overhead and can undermine the cost benefit of shutting it down.

The following approaches can help to efficiently minimise idle resources:

1. **Automate infrastructure procurement and application setup through Infrastructure as Code (IaC).** This gives the flexibility to kill and reinstate infrastructure on demand, and can even be scheduled
2. **Use low-cost cloud-based offerings for workloads that are not critical and can be restarted safely.** These are instances which are not used in cloud data centres and hence are offered at a much lower price – the catch being that they can be 'taken back' when someone else demands them.
3. **Choose reserved virtual machines (VMs) over on demand or standard VMs for critical workloads.** This is particularly applicable if they're needed for a longer duration, as they are much cheaper.
4. **Use horizontal auto scaling for VMs.** On-demand or standard VMs can be used to cover off the minimum capacity needed, with the use of spot [8,9] or preemptible[10] VM instances scaled up if the application is fault tolerant. This can be very cost effective if the application is expected to serve tens of thousands of users in peak hours, as having hundreds of spot VMs is much cheaper than maintaining the equivalent number of standard VMs.
5. **Choose a serverless approach where appropriate.** The cost of serverless resources is always based on active usage, with billing done at a highly granular level, but it is important to avoid unwanted executions[7] to keep the spending in control. For an instance exposing a serverless function to the open world would cost a lot as there will be huge number of unwanted executions.

OVERPROVISIONED RESOURCES[11]

This is applicable to IaaS and to some extent PaaS where capacity planning is needed. If the assessment of required capacity per workload is not conducted, it will result in the provisioning of resources at a higher capacity than needed, resulting in wastage due to unused capacity[12].

This can be minimized by:

1. **Provisioning for peak,** which involves applying a just-in-time, pay what you need mindset. Using auto horizontal scaling instead of large capacity procurement is a good example of this.
2. **Predetermining capacity where possible. T**his can give a good indication of how much capacity is required if autoscaling is not an option.
3. **Measure the capacity used in real-time.** Where capacity cannot be pre-determined, then there are tools (CloudWatch from AWS or Grafana, an open-source alternative) that can help highlight the capacity needed for a workload practically and make continuous adjustments as required. This creates cost visibility and enables usage forecasting.
4. For databases, provision with the **minimum required storage** capacity and use a **storage autoscaling feature** to scale storage automatically[13].

POORLY MANAGED STORAGE

Organizations generate a huge amount of data and need to maintain an accurate historical data record for audit and other purposes. Hence the cost of data is constantly increasing in parallel with its volume. To store data efficiently, it is important to classify it per the following parameters and define the storage system accordingly:

- Access requirements – how often does the data need to be read?
- Retention requirements – how long does the data needs to be kept?

Geolocation requirements – from where will the data be accessed? Here are how the above parameters can help to make the right decision on the procurement of resources:

1. **Frequency of access.** This will help determine the right storage solution. For example, Amazon S3 has several data storage tiers, and each has a specific billing rate and is suitable for various possible data access needs. It is also possible to move data between these tiers driven by rules – for example, if a file has not been accessed for 30 days, it can automatically be moved to the low-cost storage tier (e.g. Lifecycle Policies from AWS S3 and GCP).

2. **Clear retention requirements.** These can help in ensuring efficient housekeeping policies and can be configured in the cloud storage solution itself for automatic execution, reducing any manual overhead.

3. **Defining geolocation requirements,** This helps in selecting the right storage solution. For example, there is no need to obtain a multi-region cloud storage if you know the data is only going to be accessed from a specific country or region.

Also, it's important to make sure obsolete snapshots of databases, unused or obsolete VM images are removed. They are often large and add significant cost which can be avoided.

IMPROPER NETWORK TOPOLOGY

Cloud providers charge for network traffic exiting their internal network (egress) and some also apply charges when data flows across availability zones or regions. Hence, it is important to get the network topology right and have a suitable network segmentation.

Here are a few areas[14] to consider when configuring the network layer and how applications communicate through the network in cloud:

1. **Frequency Using external Ips.** Applications or workloads should use internal/private IPs or DNS hostnames when communicating with other workloads. Using the public IP for a resource in the same subnet/zone would force egress, adding an unnecessary cost overhead. This should be done whilst considering any impact on availability.

2. **Avoid cross-region data synchronization if an application is used only within one region.** This will mitigate cross-region traffic, which is more costly.

3. **Dedicated Interconnect from GCP or Direct Connect from AWS.** These should be considered for transferring large volumes of data. They are direct physical connections between an on-premises network and a cloud network, enabling cheaper and more secure data transfers when compared to purchasing additional bandwidth over the public internet.

4. **Take the application closer to its users to reduce network traffic.** For example, using CloudFront from AWS or Cloud CDN from Google can assist in serving frequently accessed content to users efficiently.

REDUCING THE RISK OF CLOUD WASTE

A robust cloud governance strategy – maintaining control of the cloud environment through the creation of rules, policies and iterative processes driven by data – is vital to managing cloud waste. Governance should be applied throughout the Software Development Life Cycle process to ensure right controls are in place for every phase.

- **Architecting cost-effective solutions** is the key. In the past organizations hosted IT products and solutions from a finite set of already procured infrastructure and resources provisioned for peak load (On premise data centres), and IT product architects were free to use these resources as they are already procured. Cloud has reversed this paradigm, as it allows for more precise solutions to meet the requirements of a specific workload while attaching a price tag to each component used in the solution.
- **Establishing Cloud FinOps Team**. This is a collaborative, data-driven way of managing cloud spend and allows finance, technology and business to drive financial accountability and accelerate the realization of business value throughout a cloud transformation.[15]
- **Iterative activities to keep cloud spend in control.** The number of iterations can be based on the number of changes done in the infrastructure:
    - **Optimize the resources** (right sizing, right features) – asses the current cloud estate to find resources that can be optimized
    - **Observability –** we cannot manage something that we cannot measure, so use observability tools to get a transparent and continuous view of cost
    - **Effective governance –** will ensure best practices are applied to ensure cloud resources are used in a cost-effective way.

As Eric Lam[16] (Global Cloud FinOps, Google having 17 years of experience with Fortune 100 customers. Recognised for helping clients create sustainable business outcomes, drive financial accountability and accelerate value realization through digital transformation) notes, “organizations who are living in cloud or planning to move to cloud often encounter a set of common problems”. These include understanding the cloud’s billing structure and optimizing its cost; accountability and transparency issues on infrastructure procured and its cost; and siloed and partial optimizations which are not centrally managed.

It is important to have in-house capabilities to address these challenges from strategic point of view. Subject matter experts can also help organizations establish governance and policies to manage cloud spend, realize business value and effectively set up a cloud FinOps process within the organization.

IN SUMMARY

Cloud waste management is not just about cutting cost – it is also about knowing where to spend money to maximize business value. This requires an iterative and continuous process that provides a consistent methodology to visualize and manage cloud consumption in the most cost-effective way.

Whether you are starting your cloud adoption journey or you have already onboarded to cloud, Capco will be pleased to share our experiences on cloud waste management.

REFERENCES

1. **Moore, Susan**. Gartner Says More Than Half of Enterprise IT Spending in Key Market Segments Will Shift to the Cloud by 2025. gartner.com. [Online] February 9, 2022. https://www.gartner.com/en/newsroom/press-releases/2022-02-09-gartner-says-more-than-half-of-enterprise-it-spending.

2. **Flexera.** CM-REPORT-State-of-the-Cloud. Flexera. [Online] https://info.flexera.com/CM-REPORT-State-of-the-Cloud.

3. **McKendrick, Joe.** One-Third Of Cloud Spending Wasted, But Still Accelerates. www.forbes.com. [Online] April 29, 2020. https://www.forbes.com/sites/joemckendrick/2020/04/29/one-third-of-cloud-spending-wasted-but-still-accelerates/?sh=7629119c489e.

4. **Gartner.** Gartner Forecasts Worldwide Public Cloud End-User Spending to Reach Nearly $500 Billion in 2022. gartner.com. [Online] April 19, 2022. https://www.gartner.com/en/newsroom/press-releases/2022-04-19-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-500-billion-in-2022.

5. **Microsoft.** Shared responsibility in the cloud. docs.microsoft.com. [Online] March 01, 2022. https://docs.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility.

6. **Amazon.** Optimize and Save your IT costs. aws.amazon.com. [Online] https://aws.amazon.com/aws-cost-management/aws-cost-optimization/.

7. **Google.** Optimize cost: Compute, containers, and serverless. cloud.google.com. [Online] April 06, 2022. https://cloud.google.com/architecture/framework/cost-optimization/compute.

8. **AWS, Amazon.** Amazon EC2 Spot Instances. aws.amazon.com. [Online] https://aws.amazon.com/ec2/spot/.

9. **Azure, MS.** Azure Spot Virtual Machines. azure.microsoft.com. [Online] https://azure.microsoft.com/en-gb/services/virtual-machines/spot/.

10. **Google.** Preemptible VM instances. cloud.google.com. [Online] May 04, 2022. https://cloud.google.com/compute/docs/instances/preemptible.

11. **Fortney, Elaina**. 6 Types of Overprovisioned Resources Wasting Money on Your Cloud Bill. www.parkmycloud.com. [Online] July 5, 2018. https://www.parkmycloud.com/blog/overprovisioned-resources/.

12. **Liu, Will.** The Cloud Waste Problem: How to Stop Overprovisioning Resources in 2022. cast.ai. [Online] February 22, 2022. https://cast.ai/blog/the-cloud-waste-problem-how-to-stop-overprovisioning-resources/.

13. **Google.** Reduce overprovisioned Cloud SQL instances. cloud.google.com. [Online] April 06, 2022. https://cloud.google.com/sql/docs/mysql/recommender-sql-overprovisioned.

14. —. Optimize cost: Networking. cloud.google.com. [Online] April 06, 2022. https://cloud.google.com/architecture/framework/cost-optimization/networking.

15. —. Google Cloud FinOps. cloud.google.com. [Online] https://cloud.google.com/learn/what-is-finops.

16. **Eric Lam, Pathik Sharma.** Cloud FinOps: The Secret To Unlocking The Economic Potential Of Public Cloud. www.forbes.com. [Online] April 04, 2022. https://www.forbes.com/sites/googlecloud/2022/04/04/cloud-finops-the-secret-to-unlocking-the-economic-potential-of-public-cloud/.

AUTHOR

Jyotirmoy Mukherjee
Principal Consultant and Solutions Architect

CONTACT

Peter Kennedy, Partner and Cloud Lead, peter.kennedy@capco.com

ABOUT CAPCO

Capco, a Wipro company, is a global technology and management consultancy specializing in driving digital transformation in the financial services industry. With a growing client portfolio comprising of over 100 global organizations, Capco operates at the intersection of business and technology by combining innovative thinking with unrivalled industry knowledge to deliver end-to-end data-driven solutions and fast-track digital initiatives for banking and payments, capital markets, wealth and asset management, insurance, and the energy sector. Capco's cutting-edge ingenuity is brought to life through its Innovation Labs and award-winning Be Yourself At Work culture and diverse talent.

To learn more, visit www.capco.com or follow us on Twitter, Facebook, YouTube, LinkedIn Instagram, and Xing.

WORLDWIDE OFFICES

APAC
Bangalore
Bangkok
Gurgaon
Hong Kong
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE
Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA
São Paulo

WWW.CAPCO.COM

CAPCO
a wipro company