Spot
by NetApp

Guide

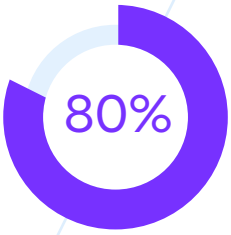
The FinOps guide for cloud excellence

Introduction

What is FinOps and why is it so important

The rapidly emerging discipline of FinOps is defined as the "practice of bringing financial accountability to the variable spend model of cloud, enabling distributed teams to make business trade-offs between speed, cost, and quality[1]." While theoretically always important, FinOps today is positioned to play a critical role in the success of every organization using the cloud.

Surveys indicate that as many of 40% of IT decision makers cite cost savings as their primary motivation for moving to the cloud. However, according to a study by 451 Research, 53% of organizations that decided to move to the cloud indicate that cost is still a significant pain point[2]. Further, Gartner predicts that through 2020, 80% of organizations will overshoot their cloud IaaS budgets due to a lack of cost optimization approaches[3].



Gartner predicts:

80% of organizations will overshoot their cloud IaaS budget!

Controlling and reducing cloud costs has become more urgent than ever before in the volatile environment caused by pandemic outbreaks and measures taken in response to them. The assumptions and projections that companies had previously been using for demand, sales, revenue and income have suddenly become historical relics. Companies are forced to reckon with sudden and unpredictable changes in demand while also recalibrating their budgets. Reining in cloud costs is not only critical to bringing costs and revenues into alignment, it's also critical to making it possible to retain key resources that will be needed in the future.

Faced with these pressures, organizations are looking to establish FinOps teams to drive the change necessary for financial accountability, operational efficiency and overall excellence in the cloud. Based on our experiences working with over 1,500 organizations, this FinOps guide provides clear guidance for delivering a highly available cloud that meets the needs of all stakeholders with the greatest cost-efficiency possible.

1. https://www.finops.org/what-is-finops/
2. "Voice of the Enterprise: Cloud Transformation, Organizational Dynamics", 451 Research.
3. "How to Identify Solutions for Managing Costs in Public Cloud IaaS", 19 August 2019, Gartner, Inc.

Starting with the right framework

To create financial accountability and efficiency in the cloud, ideally the entire company should be involved. However, all of the following can be implemented within individual business units and teams and with success, expanded company-wide.

Implementing FinOps practices doesn't require a big or complex consulting project.

Here are the key steps to follow:

1 Cross-organizational alignment

As most FinOps efforts will impact multiple teams, all relevant stakeholders must understand and commit to the process. Being able to present and correlate your business KPIs (e.g. desired cost per customer, product, etc.) to related cloud spend, will make it easier to gain management buy-in and explain FinOps and cost management activities.

2 Understanding your organization's needs

Before getting started, it's imperative to get a clear view of your organizational landscape with perspective on the goals of individual business units and teams. The following activities are a good place to begin:

- Interviewing engineering teams to understand their current, near-term and future projects to anticipate their needs for cloud resources
- Reviewing engineering's projects with your DevOps, SREs, and TechOps teams to verify requirements for cloud architecture, associated infrastructure and budget.
- Verifying with the Finance team that the requested cloud spend is in line with available budget.

Along the way, these interviews will give you familiarity with the various projects and organizational hierarchy which will be essential for tagging all cloud resources, which itself is a prerequisite for accurate spend visibility and cost allocation.

3 Mandatory tagging and cloud spend monitoring

Once you know which tags should be applied to deployed resources, a mandatory tagging strategy combined with real governance should be established. Open source tools such as Cloud Custodian offer a comprehensive library of scripts for reporting on and enforcing compliance with fundamental aspects of cloud management such as new resource labeling or tagging.

Once proper tagging is in place, the next step is to gain comprehensive and unified reporting of ypur cloud spend - all based on the business logic of your own company. This can be done by leveraging cloud providers' native tools such as AWS Cost Explorer, Google Cloud Billing Report and Azure Cost Management. Third-party solutions such as Spot by NetApp's Cloud Analyzer offer a single pane of glass view for AWS, Azure and GCP.

4 Assess and prioritize savings opportunities

Mapping your cloud infrastructure footprint can leave you with an overwhelming amount of information. To identify the best places to reduce costs, it is important to balance the potential savings against the complexity and resources required to realize those savings. A systematic approach that assesses compute, storage, and network infrastructure is crucial. Modern tools that calculate potential savings, not just summarize current costs, are important to leverage in this process.

5 Establish ongoing processes that reduce costs today and in the future

One-time cost reductions are often the first step taken to save money, however even more important is ensuring that cloud infrastructure costs are kept under control in the future as well. Without that, inefficiency is almost certain to increase again until the next cost reduction project.

Hands-on cloud cost management

Reducing cloud compute costs

Compute infrastructure is typically one of the largest parts of an organization's cloud bill, and as a result often the source of the most significant opportunities to reduce costs.

Here are key ways to quickly reduce those costs:

1

Identify idle compute resources

The ease of provisioning instances and containers in the cloud is a double-edged sword, especially in large teams without mature governance controls. Not only is it all too easy to provision resources but then leave them running even when no longer needed, but errors in deployment templates can also deploy unneeded resources or leave them orphaned after post-job cleanup.

To identify unused resources, cloud management tools can be used that look at metrics such as network traffic, CPU load, and similar data points to identify resources that are no longer active. Even better are tools that deploy advanced analytics and automation to monitor resources continuously, and automatically shut down those that are no longer in use. Tools that provide "showback" of costs by instances, tags, pods, clusters, etc. are also valuable to expose excess costs.

2 Optimize purchasing strategies

Cloud platform vendors have introduced a wide array of usage-based pricing models, and that is particularly true when it comes to compute resources. The same cloud compute resource can have a vast array of price points, depending on factors such as region, data center zone, and pricing plan. On-demand, reserved instances, convertible reserved instances, preemptible instances, spot instances and Savings Plans are just some examples of the compute pricing options offered in public cloud infrastructure.

The savings possible from intelligent purchasing strategies can be enormous because prices across these different options can vary dramatically. For example, reserved instances can be 75% less expensive than on-demand instances, and spot instances are up to 90% less expensive than on-demand instances.

Where can you find reserved capacity?

When it comes to reserved instances and cost allocation, many companies use a centralized approach to procurement so that all business units and projects will benefit from lower pricing. However, if necessary you can buy reserved capacity on the sub-account level without sharing the benefits with other accounts.

Putting in place tools and processes that identify where reserved capacity is appropriate or where spot or preemptible instances can be leveraged, as well as continuously balancing and optimizing the mix of pricing options in use can deliver significant savings without requiring disruptive changes to infrastructure.

aws

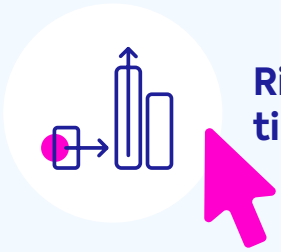
AWS Reserved Instances are also available for RDS, Redshift, Elasticsearch, Elasticache, and DynamoDB.

In Azure, RIs cover Blob storage capacity, the compute component of Azure Database for MariaDB, MySQL, and PostgreSQL, as well as quite a few other Azure services.

In GCP, Committed use discounts can be applied to VM usage only with 1 or 3 year terms providing up to 70% cost reductions.

3 Right-size your compute resources

In traditional infrastructure, capacity planning is a crucial exercise that guides purchases of hardware and related software licenses. Although resources in the cloud can be provisioned on-demand, capacity planning challenges and inflexible infrastructure still present problems in the cloud. Overprovisioning remains common, driven by a need to have buffer capacity to absorb resource failures and spikes to maintain production SLAs consistently. It's particularly true in container infrastructure, where inefficient bin packing often leads to overprovisioning.



Right-sizing tip

Sometimes right-sizing can yield **dramatic results** as seen here with the c5n.large.

It provides all the networking capacity needed for communication-intensive workloads without the excess CPU and Memory baggage that the c4.8xlarge has and for **94% cost savings!**

Instance	Price	vCPU	Memory (GiB)	Instance storage	Network performance
c4.8xlarge	$1.591 per hour	36	60	• EBS only • Dedicated EBS • Bandwidth (Mbps): 4,000	10 Gigabit
c5n.large	$0.108 per hour	2	5.25	• EBS only • EBS Bandwidth (Mbps): Up to 4,750	Up to 25 Gigabit

Leveraging monitoring and analytics tools to assess actual resource usage of instances, containers, clusters and pods can help you identify what resources you need to support workloads and predict what resources will be needed in the future, making it possible to improve container packing efficiency. That information makes it possible to confidently size resources and create adequate buffer capacity to meet workload requirements, deploying (or removing) resources at just the right time to make the most efficient use of infrastructure.

In some companies, the FinOps team will enforce right-sizing, giving teams a 2 week notice that they need to right-size resources or justify over-provisioning, before shutting said resources off.

Supersizing really does save you money!

As seen in the chart below, by going a few sizes up, we can achieve better resource allocation and **reduce the cost by 20%.**

	Pod count	Pod M	Instance M	Memory %	Pod vCPU	Instance vCPU	CPU %	Cost per pod (OD hourly)
m5.large	1	6,500	8,000	81.25%	1.5	2	75%	$0.096
m5.xlarge	2	13,000	16,000	81.25%	3	4	75%	$0.096
m5.2xlarge	5	31,500	32,000	98.44%	7.5	8	93.75%	$0.077

Example: Pod requires 6,500mb of memory and 1.5vCPUs

Reducing storage infrastructure costs

From data lakes to repositories for application state, storage is used in a wide-ranging set of scenarios in the cloud, also presenting opportunities for reducing cloud costs.

4 Release unneeded storage capacity

Because cloud storage is quick to provision and practically unlimited, it is easy to end up with unnecessary storage capacity: provisioning oversized volumes, allocating storage volumes that don't get used, having orphaned volumes or keeping snapshots that are no longer needed.

Storage management and monitoring utilities available natively from cloud platform vendors as well as from third-party providers can quickly identify orphaned volumes and snapshots for removal, whether for Amazon EBS volumes, Azure Virtual Disks or GCP Block Storage. Usage data can help to identify overprovisioned volumes that can be right-sized, while snapshot retention policies can be reviewed and modified to ensure that storage snapshots that are no longer needed are not unnecessarily retained.

5 Leverage storage tiering

The storage resources available in the cloud come in a wide range of options with different latency, throughput and cost characteristics. Storing data in the best-fit storage tier and moving data between tiers if and when appropriate can significantly reduce cloud storage costs.

For example, recently-created data in active use may require the low latency and sustained throughput provided by solid-state disks (SSDs) and provisioned throughput, while "warm" data can be stored on lower-cost spinning disks. Additional storage tiers such as object storage (e.g. Amazon S3, Azure Blob Storage or Google Cloud Storage) can be used for use cases where latency requirements are less demanding, and cold storage tiers can be used for long-term archive data. The storage lifecycle policies provided by cloud platforms can be used to automate migration of data between tiers as well.

6 Align storage redundancy with requirements

Cloud storage services such as Amazon S3, Azure Blob Storage and others offer configurable levels of redundancy, from the number of local copies of data to the number of remote replicas of data. Some applications also offer their own data redundancy within the application itself.

Using reduced redundancy configurations for non-critical data can immediately bring down storage costs. For example, cloud object storage with reduced redundancy can be up to 50% less expensive than standard object storage.

Reducing network costs

Cloud computing platforms vary significantly in their pricing for data transfer, however costs associated with data transfer in the cloud can quickly become significant. Without rearchitecting applications, these steps can help bring down those costs.

7

Reduce traffic across zones and regions

Network traffic between datacenters can be an overlooked but significant contributor to cloud costs, not only for traffic between availability zones but even more for traffic between regions. Although the need to ensure redundancy and resiliency is one reason that creates cross-datacenter traffic, the ease of provisioning services in multiple availability zones can also make it easy to inadvertently create significant amounts of unnecessary traffic between regions or between availability zones. The data transfer costs that result can quickly add up, especially for "chatty" workloads as well as for applications that consist of large numbers of distributed services.

Rebalancing services across zones to minimize data transfers across regions and zones provides one way to reduce these costs. Done thoughtfully, rebalancing can significantly reduce cross-datacenter communication without compromising resiliency, and can even help improve performance. Network tracing and logging tools can also help, bringing to light misconfigurations that are creating unnecessary cross-datacenter traffic that can be quickly reduced.

Deciding which **regions and zones** to run your workloads in requires careful consideration of governmental regulations, customer needs and of course, costs.

8 Optimize network configurations

Not only does the amount of network data transfer impact costs, so does the way in which networks are configured. Depending on how network traffic is routed, data transfer costs can vary widely and grow quickly with usage.

Modifying network configuration can have a significant impact on data transfer costs without affecting available throughput. For example, choosing private IP addresses rather than public or elastic IP addresses where possible can have a big impact on data transfer costs.

9 Deploy distribution and caching solutions

Many applications and application components make requests for the same data from other cloud services or from remote repositories. Particularly for media objects and for transfers of large data sets for data processing and analytics, the repeated data transfers created by these requests can drive up costs. For example, transfers of data from object storage services like Amazon S3 can incur costs that vary based on location and configuration.

Deploying content distribution networks and caching services can help reduce these costs by reducing repeated transfers from remote services and locations. These solutions do come with their own costs, making it important to first assess their cost in order to determine the potential savings from deploying them.

Summary

Keeping cloud costs low

Scheduling lower environment workloads to run 8AM–6PM on weekdays instead of 24/7 can net you 70% cost savings which is on par with the biggest RI discounts achievable.

It's great to take specific steps to reduce cloud costs in the face of an urgent need to fit within your budget, but all too often cost optimization is approached as a point-in-time project. However, it is even more important to bring your costs under control on an ongoing basis to avoid silently increasing inefficiencies that can drain your budget again in the future.

These best practices are critical to ensuring that you are continuously operating as efficiently as possible in the cloud:

Use machine learning and analytics for smarter cost management

There is certainly no shortage of tools that can generate charts, graphs, reports and alerts. However, just having access to more charts, graphs and reports can quickly lead to information overload. Already stretched FinOps teams can find themselves overwhelmed by an array of data to understand and alerts to investigate, too many of which turn out to be false alarms driven by natural changes and evolution in resource usage. Instead, take advantage of modern tools that use machine learning and artificial intelligence to learn resource usage patterns, making it possible for them to identify true anomalies and alert you only to changes that truly require closer examination.

Schedule resource usage

For many applications and services, utilization changes are driven by business schedules, recurring processes or business events that are known in advance. For example, cloud resource requirements can change dramatically because of end-of-quarter processing or Monday morning business activity. Scheduling scaling and resource allocation in advance based on that knowledge helps ensure not only that resources are available when needed, but can also be used to ensure that they do not continue to run idle, driving up costs, when they are no longer needed.

Leverage automation

Ensuring that these best practices are continuously applied requires automation. For example, concern about how quickly and easily scaling can be done to react to changes in demand often leads to overprovisioning. Even in the cloud, the full sequence of steps needed to bring new resources online can take a significant amount of time--there are post-deployment configurations that need to be applied to those resources, software that needs to be validated and launched in those instances, changes to network traffic management and more. Intelligent automation helps address overprovisioning costs: with the right solution, the need to scale is detected automatically and acted on based on rules configured in advance, reducing delay and the need to overprovision infrastructure.

Reducing cloud costs with Spot by NetApp

Taking steps to reduce and control cloud infrastructure costs requires ongoing effort and diligence that could all too easily consume significant amounts of time for already overstretched FinOps teams. To help address that challenge, Spot has created a suite of products built on unique machine learning and analytics that monitor cloud workloads and resources to provide visibility, guidance and automation that continuously optimize cloud infrastructure costs without compromising availability, performance and flexibility.

Spot's solutions operate in all the leading public cloud environments, ensuring optimized infrastructure for both legacy and cloud-native applications.

Learn more about Spot by NetApp's products and solutions online at spot.io

8.17dec20