



9

Checklist

Ways to reduce cloud costs now and in the future



53%

say that
cloud cost is
a significant
pain point¹

Why reducing cloud costs has become a priority

Surveys indicate that as many of 40% of IT decision makers cite cost savings as their primary motivation for moving to the cloud. However, that hope is all too often met with disappointment: 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 point¹. Further, Gartner predicts that through 2020, 80% of organizations will overshoot their cloud IaaS budgets due to a lack of cost optimization approaches².

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. That's had major implications for IT organizations as well, as they are forced to reckon with sudden and unpredictable changes in demand while also recalibrating their budgets. Reining in IT 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 urgently reexamining their cloud expenses and looking for practical ways to reduce costs without adding risk or requiring difficult, disruptive changes to existing infrastructure. Based on our experiences working with over 1,000 organizations, this checklist focuses on ways that you can reduce your cloud infrastructure costs, both now and into the future.

1. "Voice of the Enterprise: Cloud Transformation, Organizational Dynamics", 451 Research.

2. "How to Identify Solutions for Managing Costs in Public Cloud IaaS", 19 August 2019, Gartner, Inc.

Starting with the right framework

To reduce your cloud infrastructure costs and keep them low, it's important to start with the right approach. Given the complexity of cloud infrastructure, it's easy to get quickly lost in the details, diving into making labor-intensive changes to infrastructure and application design that can end up taking significant effort, but only produce modest savings.

Reducing cloud costs doesn't need to be a big or complex consulting project. The key steps to follow:

- **Discover and understand your cloud infrastructure.**

Because cloud resources and services can be provisioned in just a few clicks or API calls, it's easy to quickly lose track of what is deployed. Tools that can automatically discover and map your infrastructure can help you get that full picture, not only showing information about instance counts and summary costs across all of your cloud accounts, but also providing the ability to look at that information organized by tags, pods, clusters, services and applications.

- **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.

- **Take steps 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.

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 often up to 90% less expensive than on-demand instances. 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.

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.

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.

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.

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.

Keeping cloud costs low

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 site reliability (SRE) and CloudOps 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

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 CloudOps 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 cloud environments, ensuring optimized infrastructure for containers, Kubernetes, autoscaling applications and more.

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