EBOOK

AWS Cost Optimisation

aws | POLAR SEVEN

Introduction

AWS Cost Optimisation

Most companies who make a significant cloud investment find their AWS bill grows much faster than expected.

There are many resources in AWS which are not equivalent to on-prem solutions, so it is hard to make sure you are not paying too much. A 'Cost-Optimised' system seeks to fully utilise all resources, and meet your functional requirements for the lowest price point.

PolarSeven's experience in AWS Cloud solutions has highlighted around 65% of expenditure to be attributed to EC2, 20% to RDS and the next biggest being 5% on S3, with the remaining 10% divided into much smaller costs. In general, most savings will come from reducing cost for EC2 and RDS. How these costs are reduced will vary based on your applications, your users, and predictability of demand.

This document will explore the top 10 practices we have found most effective at reducing cost.

65%

of AWS bill spent on EC2

20%

of bill spent on RDS

5%

of Cloud bill spent on S3

Contents

> ***Rightsizing should be done whenever you have a change in usage patterns, AWS price drops, or when newer and more efficient resource types are released.***

EC2 Rightsizing

EC2 Rightsizing is the process of provisioning the lowest cost resource that still meets the technical specifications of a specific workload. Rightsizing should be done whenever you have a change in usage patterns, AWS price drops, or when newer and more efficient resource types are released.

AWS provide several tools for customers to accurately identify the usage and cost of systems, allowing dynamic provisioning of resources based on short term demand. For those who have purchased a Business or Enterprise Support Plan, AWS offer an online tool called Trusted Advisor, which highlights different optimisations applicable to your system.

Trusted Advisor highlights EC2 instances that have had less than 10% average daily utilisation on at least 4 of the previous 14 days. These instances should be evaluated to determine if they are optimally sized.

There are several recommended actions for underutilised EC2 instances depending on the findings. For example, a server that is displaying 0% utilisation for an extended period of time can most likely be terminated or stopped. Other instances may have consistently low utilisation for an extended period of time and can be changed to a more cost effective instance type.

* *Please be aware that reducing instance size can have a significant impact to the performance of your workload, and should be thoroughly tested before deploying to production.*

Trusted Advisor

Trusted Advisor is a useful tool that will **proactively recommend improvements** to reduce cost, improve performance, and increase security. For those with business or enterprise support, it will determine which instances are underutilised based on CPU and I/O metrics.

Stopping

An instance can be stopped and started as long as an EBS volume is attached as its root device. A stopped instance will not incur usage or data transfer fees; however **AWS do charge for the storage of the EBS volume** (which is considerably less).

Terminating

If you decide an instance is no longer required, it can be terminated and you will no longer be charged. It will remain visible in your AWS console a short while after termination. By default the Root EBS volume will also automatically delete, however by default any additional EBS volumes that were attached at launch or during its lifecycle will be preserved. This is important to note as these **EBS volumes will continue to incur a cost.** If left ignored, in just a few months you could find yourself with a bill that has doubled in cost simply due to this. This can be modified using the volume's **DeleteOnTermination** attribute.

Changing Instance Type

If an instance has consistently low utilisation, it may be able to **maintain performance with a lower cost instance type.** Analysis of detailed metrics in CloudWatch or third party monitoring tools from vendors such as **New Relic** will help you determine which instance type is most suitable for your workload. As new instance types are continually released, and workloads are constantly evolving, this process should be completed on a regular basis.

Savings
of up to 70%

Effective
for predictable workloads

Cost
can increase if forecast wrongly

Reserved Instances

Reserving instances can provide **savings of up to 75%** when compared with on-demand usage. Reservation is most effective for predictable workloads, as you must make assumptions on usage for a 1-3 year period.

Before deciding if reserved instances are suitable for a specific use case, ask:

1. How long have the instances been running?
2. Do you expect any major changes to the applications or load upcoming? .
3. Do the instances run 24/7?
4. Is the instance correctly sized?
5. What are the expected lifetimes of the instances?

These are important considerations when deciding whether to reserve capacity and for how long. Reservations can lead to decreased savings if not calculated correctly, and in some cases can cost more than if you went with On Demand instances. AWS provide a free reservation recommendation through their Cost Explorer, that will assist you in determining which instances are candidates for reservation.

Reserved instance **discounts can be applied to any instance within the same instance family.** This provides flexibility to move change instance sizes without losing the reserved savings. If you have a group of instances in the same family, a good practice is to reserve a portion of these so that if you want to terminate or change instance types, there will be another instance available to realise the reserved discount.

If you still want to realise reserved discount savings while having the flexibility to move to a completely different instance family, AWS provide the Convertible Reserved Instance offering. Convertible Reserved Instances are useful when workloads are likely to change; or you want to be able to take advantage of future price drops - or you are unable to do capacity planning or forecasting.

90%
Saving

ElastiCache for Read-Heavy Applications

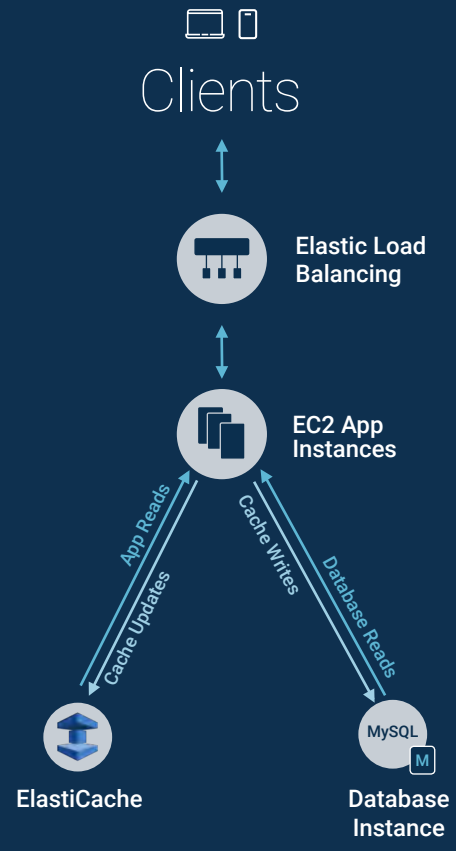
Amazon ElastiCache allows you to improve load and response times to user actions and queries, while reducing the cost associated with scaling web applications.

ElastiCache is an in-memory data store, used for read-heavy applications where users demand real-time response. It stores subsets of data from the database, forming a 'caching layer'. When a read request is sent, ElastiCache checks to see if it has the answer and returns it. If the cache does not have the required data, the application will retrieve it from the database, and the data will be stored for subsequent reads by ElastiCache.

Processing reads with AWS ElastiCache is much more efficient, as a single node of in-memory cache can deliver the same read throughput as several database nodes. The caching layer therefore saves significant money as you're only paying for one node, instead of multiple database nodes.

Let's have a look at how much you could save.

In this scenario, you could be saving over 90% by using ElastiCache over the conventional MySQL Database instance. Amazon ElastiCache delivers the high performance and throughput benefits organisations seek from an in-memory data store. One node of ElastiCache can offer hundreds of thousands of calls - sometimes up to a million calls per second: that's one to two times more than a disk-based database. The latency for a call to ElastiCache can be 300-500 microseconds compared to double-digit milliseconds for a traditional database.



Example: **30,000 reads per second**

db.m3.large
30,000 PIOPS - Single-AZ: $3,889.68

cache.m3.large
30,000 GETS per second: $355.02

Balance availability and cost with AWS Auto Scaling

AWS Auto Scaling allows your application to restart if it has an error, or dynamically increase the number of application instances to keep your service running even while under DDOS attack. Auto Scaling also enables you to increase or decrease capacity based on time, or if certain usage thresholds are breached.

> *Please be aware that reducing instance size can have significant impact to the performance of your workload, and should be thoroughly tested before deploying to production.*

Auto Scaling can be used in conjunction with AWS CloudWatch, which can send alarms to trigger specific scaling down activities when CPU utilisation or the request count is low.

If there is no requirement for people to access certain resources after hours, Auto Scaling can schedule these resources to stop and start at the necessary times. The maths follows; if your business operates 9am-5pm, Mon-Fri, by turning off resources outside these hours, you would save:

168	-	40	=	118	This would give 75% savings compared to running the instances 24/7.
hours		**hours**		**hours**	

> ***By turning off resources outside of working hours, you can achieve 75% savings compared to running the instances 24/7.***

Purchase Spot Instances

Amazon EC2 Spot instances allow customers to purchase unused EC2 instances in the AWS cloud at a significantly lower rate by sacrificing availability. Similar to On-Demand instances, there is no upfront commitment when purchasing a Spot instance, you pay only for the compute consumed while the instance is running. The main difference seen between On-Demand and Spot instances is that Spot instances can be interrupted with as little as 2 minutes notification that EC2 requiresthe capacity back. This can be due to AWS utilising all EC2 instances, or the Spot price raising higher than your bid.

This positions it to be the most effective for stateless or flexible applications, such as large batch processing of data, continuous integration/ continuous development and other dev/test workload scenarios. It allows customers to take advantage of the operating scale seen within AWS and optimise their workload costs by supplementing On-Demand and reserved instances.

Use Spot Instances for **stateless apps,** or for flexible workloads

Save, by **sacrificing** availability

Participate in the world's largest Cloud **Compute market!**

Run on Serverless with AWS Lambda

How Lambda is charged

Lambda functions are a new way in which customers can think about compute resources. Rather than paying for an instance whenever it is in an available state - Lambda only runs when triggered. Amazon then calculates the cost of Lambda based on the number of times it is triggered, the duration it is running, and the amount of memory provisioned. Each time a Lambda is triggered, the duration & memory are multiplied, and your consumption is measured in Gigabyte-seconds: you will see the unit 'GB-sec' on your Amazon bill.

Optimising Allocated Memory

As allocated memory is one of the variables Amazon uses to charge Lambda, it makes sense to focus your attention on reducing this. Often, Lambda functions uses much less memory than is allocated. CloudWatch (an out of the box service provided by Amazon) can be monitored to determine how much memory your functions are actually utilising. Once you have this information, you can incrementally decrease the allocated memory, and ensure that the duration to run the Lambda has not increased as a result.

Optimising Lambda Duration

Considering the minimum billable duration of a Lambda function is 100ms, any function that runs for less than this will still be charged at the same rate. As you will be charged for the same amount - why not reduce the allocated memory to see what happens? You would expect by decreasing the memory the Lambda would take longer to run, but due to the minimum charge duration, you will pay less as result of decreasing memory.

> ***"Serverless does not equal 'No servers', but rather a change in the way customers think about compute resources.***

Build infrastructure according to AWS best practices

Adopt a flexible consumption model

The elasticity experienced from using AWS allows businesses to dynamically provision and retire resources based on either real-time data, or predictable business patterns. For example, development and test environments are often only used during business hours. Resources can be automatically provisioned to run specifically during these hours for savings of 75% (40 hours compared with 168). Similarly, environments can be dynamically provisioned when there is an increase or decrease in load.

Continually analyse efficiency

AWS CloudWatch provides in-depth analytics for each resource, application, and service, allowing the business to clearly correlate costs of workloads with the cost to deliver it. Consistent analysis of resource utilisation allows the business to continually optimise costs based on usage.

Automation

A major risk posed in many IT architectures is that of human error, or accidental configuration change. AWS helps alleviate this through the implementation of Automation. The most common area to focus on is your deployment workflow, as this is the point at which changes are introduced, and failure or disruption of services occurs more frequently. You should also find a reduction in effort & time spent to deploy new features, eventuating in a reduction in operational expense. Automation has the potential to reduce overall business costs, reducing salary/ workforce spend, and lowering the frequency of downtime.

Engage with the Experts

Amazon has a wide network of partners with a variety of different skill sets who have gone through the joy and pains of migrating to the Cloud and exploring new technologies. Your time is best spent focused on your own business: for niche, highly specialised work, engage specialists who spend all their time perfecting the art of AWS DevOps.

Audit your AWS Accounts

The cloud economy is constantly changing. Hundreds of new AWS features & services are released each year, and customers need to keep up with the fast pace in which the industry moves. For this reason, a big part of keeping AWS costs at bay is to regularly audit your Amazon accounts, to uncover underutilised or over-specced resources.

Luckily, AWS provides some 'out of the box' services that will automatically scan and monitor your environment and highlight potential cost savings. The most common of these is Trusted Advisor, an easy to use tool that will suggest areas to remediate based on the 5 pillars of a well-architected system - one of which is Cost Optimisation.

A recent customer we engaged with uncovered months of unused EBS volumes - stemming from one checkbox that was not checked to automatically delete. After a mere 3 months, this customer's AWS bill had increased by 40%, despite application usage remaining stable.

For the AWS accounts that we support at PolarSeven, we conduct a monthly audit with respect to cost, security, performance & operations. This consistent review ensures that our customers do not experience bill shock at the end of each month.

Leverage AWS Credits

AWS credits are a great way for you to save on your monthly AWS service bill. Credits can be awarded for a variety of activities such as you or your team completing an AWS certification course; attending webinars & events; conducting academic research activities; or even purchasing certain Marketplace solutions, to name a few.

Credits are applied to your AWS account, and are consumed by eligible services until their value is depleted.

If you are a startup...

The AWS Activate program was created to allow Startups to focus on building and scaling their environment with up to $100,000 of AWS service credits. This program is available to customers in all industries, whether it be location services, education or travel.

Your Startup may be eligible for up to $100,000 of AWS credits, under the AWS Activate program

Choose a managed service that includes cost optimisation

PolarSeven specialises in AWS Managed Services, with a service that includes cost optimisation built in.

Leverage our expertise

We understand which AWS products and purchasing methods to leverage to best control and manage cloud costs, now and into the future. Don’t rely on just your internal know-how to get it right.

Automate your infrastructure with us

We are all AWS DevOps professionals at PolarSeven. Most of our customers come to us for our automation-powered professional services, and then retain our services ongoing to augment their internal team.

Allow us to audit your bill

We proactively look for savings in your AWS bill, month-in, month-out without fail.

Level 13, 135 King Street, Sydney, 2000
1300 659 575 | hello@polarseven.com

www.polarseven.com

aws partner network