# Cost Optimization Pillar

AWS Well-Architected Framework

*April 2020*

# Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

# Contents

# Abstract

This whitepaper focuses on the cost optimization pillar of the Amazon Web Services (AWS) Well-Architected Framework. It provides guidance to help customers apply best practices in the design, delivery, and maintenance of AWS environments.

A cost-optimized workload fully utilizes all resources, achieves an outcome at the lowest possible price point, and meets your functional requirements. This whitepaper provides in-depth guidance for building capability within your organization, designing your workload, selecting your services, configuring and operating the services, and applying cost optimization techniques.

# Introduction

The AWS Well-Architected Framework helps you understand the decisions you make while building workloads on AWS. The Framework provides architectural best practices for designing and operating reliable, secure, efficient, and cost-effective workloads in the cloud. It demonstrates a way to consistently measure your architectures against best practices and identify areas for improvement. We believe that having well-architected workloads greatly increases the likelihood of business success.

The framework is based on five pillars:

- Operational Excellence
- Security
- Reliability
- Performance Efficiency
- Cost Optimization

This paper focuses on the cost optimization pillar, and how to architect workloads with the most effective use of services and resources, to achieve business outcomes at the lowest price point.

You'll learn how to apply the best practices of the cost optimization pillar within your organization. Cost optimization can be challenging in traditional on-premises solutions because you must predict future capacity and business needs while navigating complex procurement processes. Adopting the practices in this paper will help your organization achieve the following goals:

- Practice Cloud Financial Management
- Expenditure and usage awareness
- Cost effective resources
- Manage demand and supply resources
- Optimize over time

This paper is intended for those in technology and finance roles, such as chief technology officers (CTOs), chief financial officers (CFOs), architects, developers, financial controllers, financial planners, business analysts, and operations team

members. This paper does not provide implementation details or architectural patterns, however, it does include references to appropriate resources.

# Cost Optimization

Cost optimization is a continual process of refinement and improvement over the span of a workload's lifecycle. The practices in this paper help you build and operate cost-aware workloads that achieve business outcomes while minimizing costs and allowing your organization to maximize its return on investment.

## Design Principles

Consider the following design principles for cost optimization:

**Implement cloud financial management:** To achieve financial success and accelerate business value realization in the cloud, you must invest in Cloud Financial Management. Your organization must dedicate the necessary time and resources for building capability in this new domain of technology and usage management. Similar to your Security or Operations capability, you need to build capability through knowledge building, programs, resources, and processes to help you become a cost efficient organization..

**Adopt a consumption model:** Pay only for the computing resources you consume, and increase or decrease usage depending on business requirements. For example, development and test environments are typically only used for eight hours a day during the work week. You can stop these resources when they're not in use for a potential cost savings of 75% (40 hours versus 168 hours).

**Measure overall efficiency:** Measure the business output of the workload and the costs associated with delivery. Use this data to understand the gains you make from increasing output, increasing functionality, and reducing cost.

**Stop spending money on undifferentiated heavy lifting:** AWS does the heavy lifting of data center operations like racking, stacking, and powering servers. It also removes the operational burden of managing operating systems and applications with managed services. This allows you to focus on your customers and business projects rather than on IT infrastructure.

**Analyze and attribute expenditure:** The cloud makes it easier to accurately identify the cost and usage of workloads, which then allows transparent attribution of IT costs to revenue streams and individual workload owners. This helps measure return on

investment (ROI) and gives workload owners an opportunity to optimize their resources and reduce costs.

## Definition

There are five focus areas for cost optimization in the cloud:

- Practice Cloud Financial Management
- Expenditure and usage awareness
- Cost-effective resources
- Manage demand and supplying resources
- Optimize over time

Similar to the other pillars within the Well-Architected Framework, there are trade-offs to consider for cost optimization. For example, whether to optimize for speed-to-market, or for cost. In some cases, it's best to optimize for speed—going to market quickly, shipping new features, or meeting a deadline—rather than investing in upfront cost optimization.

Design decisions are sometimes directed by haste rather than data, and the temptation always exists to overcompensate, rather than spend time benchmarking for the most cost-optimal deployment. Overcompensation can lead to over-provisioned and under-optimized deployments. However, it may be a reasonable choice if you must "lift and shift" resources from your on-premises environment to the cloud and then optimize afterwards.

Investing the right amount of effort in a cost optimization strategy up front allows you to realize the economic benefits of the cloud more readily by ensuring a consistent adherence to best practices and avoiding unnecessary over provisioning. The following sections provide techniques and best practices for the initial and ongoing implementation of Cloud Financial Management and cost optimization for your workloads.

# Practice Cloud Financial Management

Cloud Financial Management (CFM) enables organizations to realize business value and financial success as they optimize their cost and usage and scale on AWS.

The following are Cloud Financial Management best practices:

- Functional ownership
- Finance and technology partnership
- Cloud budgets and forecasts
- Cost-aware processes
- Cost-aware culture
- Quantify business value delivered through cost optimization

# Functional Ownership

**Establish a cost optimization function:** This function is responsible for establishing and maintaining a culture of cost awareness. It can be an existing individual, a team within your organization, or a new team of key finance, technology and organization stakeholders from across the organization.

The function (individual or team) prioritizes and spends the required percentage of their time on cost management and cost optimization activities. For a small organization, the function might spend a smaller percentage of time compared to a full-time function for a larger enterprise.

The function require a multi-disciplined approach, with capabilities in project management, data science, financial analysis, and software/infrastructure development. The function is can improve efficiencies of workloads by executing cost optimizations (centralized approach), influencing technology teams to execute optimizations (decentralized), or a combination of both (hybrid). The function may be measured against their ability to execute and deliver against cost optimization goals (for example, workload efficiency metrics).

You must secure executive sponsorship for this function. The sponsor is regarded as champion for cost efficient cloud consumption, and provides escalation support for the function to ensure that cost optimization activities are treated with the level of priority defined by the organization. Together, the sponsor and function ensure that your organization consumes the cloud efficiently and continue to deliver business value.

# Finance and Technology Partnership

**Establish a partnership between finance and technology:** Technology teams innovate faster in the cloud due to shortened approval, procurement, and infrastructure deployment cycles. This can be an adjustment for finance organizations previously used to executing time-consuming and resource-intensive processes for procuring and deploying capital in data center and on-premises environments, and cost allocation only at project approval.

Establish a partnership between key finance and technology stakeholders to create a shared understanding of organizational goals and develop mechanisms to succeed financially in the variable spend model of cloud computing. Relevant teams within your organization must be involved in cost and usage discussions at all stages of your cloud journey, including:

- **Financial leads:** CFOs, financial controllers, financial planners, business analysts, procurement, sourcing, and accounts payable must understand the cloud model of consumption, purchasing options, and the monthly invoicing process. Due to the fundamental differences between the cloud (such as the rate of change in usage, pay as you go pricing, tiered pricing, pricing models, and detailed billing and usage information) compared to on-premises operation, it is essential that the finance organization understands how cloud usage can impact business aspects including procurement processes, incentive tracking, cost allocation and financial statements.
- **Technology leads:** Technology leads (including product and application owners) must be aware of the financial requirements (for example, budget constraints) as well as business requirements (for example, service level agreements). This allows the workload to be implemented to achieve the desired goals of the organization.

The partnership of finance and technology provides the following benefits:

- Finance and technology teams have near real-time visibility into cost and usage.
- Finance and technology teams establish a standard operating procedure to handle cloud spend variance.
- Finance stakeholders act as strategic advisors with respect to how capital is used to purchase commitment discounts (for example, Reserved Instances or AWS Savings Plans), and how the cloud is used to grow the organization.

- Existing accounts payable and procurement processes are used with the cloud.
- Finance and technology teams collaborate on forecasting future AWS cost and usage to align/build organizational budgets.
- Better cross-organizational communication through a shared language, and common understanding of financial concepts.

Additional stakeholders within your organization that should be involved in cost and usage discussions include:

- **Business unit owners:** Business unit owners must understand the cloud business model so that they can provide direction to both the business units and the entire company. This cloud knowledge is critical when there is a need to forecast growth and workload usage, and when assessing longer-term purchasing options, such as Reserved Instances or Savings Plans.
- **Third parties:** If your organization uses third parties (for example, consultants or tools), ensure that they are aligned to your financial goals and can demonstrate both alignment through their engagement models and a return on investment (ROI). Typically, third parties will contribute to reporting and analysis of any workloads that they manage, and they will provide cost analysis of any workloads that they design.

## Cloud Budgets and Forecasts

**Establish cloud budgets and forecasts:** Customers use the cloud for efficiency, speed and agility, which creates a highly variable amount of cost and usage. Costs can decrease with increases in workload efficiency, or as new workloads and features are deployed. Or, workloads will scale to serve more of your customers, which increases cloud usage and costs. Existing organizational budgeting processes must be modified to incorporate this variability.

Adjust existing budgeting and forecasting processes to become more dynamic using either a trend based algorithm (using historical costs as inputs), or using business driver based algorithms (for example, new product launches or regional expansion), or a combination of both trend and business drivers.

You can use AWS Cost Explorer to forecast daily (up to 3 months) or monthly (up to 12 months) cloud costs based on machine learning algorithms applied to your historical costs (trend based).

# Cost-Aware Processes

**Implement cost awareness in your organizational processes**: Cost awareness must be implemented in new and existing organizational processes. It is recommended to re-use and modify existing processes where possible—this minimizes the impact to agility and velocity. The following recommendations will help implement cost awareness in your workload:

- Ensure that change management includes a cost measurement to quantify the financial impact of your changes. This helps pro-actively address cost-related concerns and highlight cost savings.
- Ensure that cost optimization is a core component of your operating capabilities. For example, you can leverage existing incident management processes to investigate and identify root cause for cost and usage anomalies (cost overages).
- Accelerate cost savings and business value realization through automation or tooling. When thinking about the cost of implementing, frame the conversation to include an ROI component to justify the investment of time or money.
- Extend existing training and development programs to include cost aware training throughout your organization. It is recommended that this includes continuous training and certification. This will build an organization that is capable of self-managing cost and usage.

**Report and notify on cost and usage optimization:** You must regularly report on cost and usage optimization within your organization. You can implement dedicated sessions to cost optimization, or include cost optimization in your regular operational reporting cycles for your workloads. AWS Cost Explorer provides dashboards and reports. You can track your progress of cost and usage against configured budgets with AWS Budgets Reports.

You can also use Amazon QuickSight with Cost and Usage Report (CUR) data, to provide highly customized reporting with more granular data.

Implement notifications on cost and usage to ensure that changes in cost and usage can be acted upon quickly. AWS Budgets allows you to provide notifications against targets. We recommend configuring notifications on both increases and decreases, and in both cost and usage for workloads.

**Monitor cost and usage proactively:** It is recommended to monitor cost and usage proactively within your organization, not just when there are exceptions or anomalies.

Highly visible dashboards throughout your office or work environment ensure that key people have access to the information they need, and indicate the organization's focus on cost optimization. Visible dashboards enable you to actively promote successful outcomes and implement them throughout your organization.

## Cost-Aware Culture

**Create a cost aware culture:** Implement changes or programs across your organization to create a cost aware culture. It is recommended to start small, then as your capabilities increase and your organization's use of the cloud increases, implement large and wide ranging programs.

A cost aware culture allows you to scale cost optimization and cloud financial management through best practices that are performed in an organic and decentralized manner across your organization. This creates high levels of capability across your organization with minimal effort, compared to a strict top-down, centralized approach.

Small changes in culture can have large impacts on the efficiency of your current and future workloads. Examples of this include:

- Gamifying cost and usage across your organization. This can be done through a publicly visible dashboard, or a report that compares normalized costs and usage across teams (for example, cost per workload, cost per transaction).
- Recognizing cost efficiency. Reward voluntary or unsolicited cost optimization accomplishments publicly or privately, and learn from mistakes to avoid repeating them in the future.
- Create top-down organizational requirements for workloads to run at pre-defined budgets.

**Keep up to date with new service releases:** You may be able to implement new AWS services and features to increase cost efficiency in your workload. Regularly review the AWS News Blog, the AWS Cost Management blog, and What's New with AWS for information on new service and feature releases.

## Quantify Business Value Delivered Through Cost Optimization

**Quantify business value from cost optimization:** In addition to reporting savings from cost optimization, it is recommended that you quantify the additional value

delivered. Cost optimization benefits are typically quantified in terms of lower costs per business outcome. For example, you can quantify On-Demand Amazon Elastic Compute Cloud (Amazon EC2) cost savings when you purchase Savings Plans, which reduce cost and maintain workload output levels. You can quantify cost reductions in AWS spending when idle Amazon EC2 instances are terminated, or unattached Amazon Elastic Block Store (Amazon EBS) volumes are deleted.

Quantifying business value from cost optimization allows you to understand the entire set of benefits to your organization. Because cost optimization is a necessary investment, quantifying business value allows you to explain the return on investment to stakeholders. Quantifying business value can help you gain more buy-in from stakeholders on future cost optimization investments, and provides a framework to measure the outcomes for your organization's cost optimization activities.

The benefits from cost optimization, however, go above and beyond cost reduction or avoidance. Consider capturing additional data to measure efficiency improvements and business value. Examples of improvement include:

- **Executing cost optimization best practices:** For example, resource lifecycle management reduces infrastructure and operational costs and creates time and unexpected budget for experimentation. This increases organization agility and uncovers new opportunities for revenue generation.
- **Implementing automation:** For example, Auto Scaling, which ensures elasticity at minimal effort, and increases staff productivity by eliminating manual capacity planning work. For more details on operational resiliency, refer to the Well-Architected Reliability Pillar whitepaper.
- **Forecasting future AWS costs:** Forecasting enables finance stakeholders to set expectations with other internal and external organization stakeholders, and helps improve your organization's financial predictability. AWS Cost Explorer can be used to perform forecasting for your cost and usage.

## Resources

Refer to the following resources to learn more about AWS best practices for budgeting and forecasting cloud spend.

- Reporting your budget metrics with budget reports
- Forecasting with AWS Cost Explorer
- AWS Training

- AWS Certification
- AWS Cloud Management Tools partners

# Expenditure and Usage Awareness

Understanding your organization's costs and drivers is critical for managing your cost and usage effectively, and identifying cost-reduction opportunities. Organizations typically operate multiple workloads run by multiple teams. These teams can be in different organization units, each with its own revenue stream. The capability to attribute resource costs to the workloads, individual organization, or product owners drives efficient usage behavior and helps reduce waste. Accurate cost and usage monitoring allows you to understand how profitable organization units and products are, and allows you to make more informed decisions about where to allocate resources within your organization. Awareness of usage at all levels in the organization is key to driving change, as change in usage drives changes in cost.

Consider taking a multi-faceted approach to becoming aware of your usage and expenditures. Your team must gather data, analyze, and then report. Key factors to consider include:

- Governance
- Monitoring cost and usage
- Decommissioning

## Governance

In order to manage your costs in the cloud, you must manage your usage through the governance areas below:

**Develop Organizational Policies:** The first step in performing governance is to use your organization's requirements to develop policies for your cloud usage. These policies define how your organization uses the cloud and how resources are managed. Policies should cover all aspects of resources and workloads that relate to cost or usage, including creation, modification, and decommission over the resource's lifetime.

Policies should be simple so that they are easily understood and can be implemented effectively throughout the organization. Start with broad, high-level policies, such as which geographic Region usage is allowed in, or times of the day that resources should be running. Gradually refine the policies for the various organizational units and

workloads. Common policies include which services and features can be used (for example, lower performance storage in test/development environments), and which types of resources can be used by different groups (for example, the largest size of resource in a development account is medium).

**Develop goals and targets:** Develop cost and usage goals and targets for your organization. Goals provide guidance and direction to your organization on expected outcomes. Targets provide specific measurable outcomes to be achieved. An example of a goal is: platform usage should increase significantly, with only a minor (non-linear) increase in cost. An example target is: a 20% increase in platform usage, with less than a 5% increase in costs. Another common goal is that workloads need to be more efficient every 6 months. The accompanying target would be that the cost per output of the workload needs to decrease by 5% every 6 months.

A common goal for cloud workloads is to increase workload efficiency, which is to decrease the cost per business outcome of the workload over time. It is recommended to implement this goal for all workloads, and also set a target such as a 5% increase in efficiency every 6-12 months. This can be achieved in the cloud through building capability in cost optimization, and through the release of new services and service features.

**Account structure:** AWS has a one-parent-to-many-children account structure that is commonly known as a master (the parent, formerly payer) account-member (the child, formerly linked) account. A best practice is to always have at least one master with one member account, regardless of your organization size or usage. All workload resources should reside only within member accounts.

There is no one-size-fits-all answer for how many AWS accounts you should have. Assess your current and future operational and cost models to ensure that the structure of your AWS accounts reflects your organization's goals. Some companies create multiple AWS accounts for business reasons, for example:

- Administrative and/or fiscal and billing isolation is required between organization units, cost centers, or specific workloads.
- AWS service limits are set to be specific to particular workloads.
- There is a requirement for isolation and separation between workloads and resources.

Within AWS Organizations, consolidated billing creates the construct between one or more member accounts and the master account. Member accounts allow you to isolate

and distinguish your cost and usage by groups. A common practice is to have separate member accounts for each organization unit (such as finance, marketing, and sales), or for each environment lifecycle (such as development, testing and production), or for each workload (workload a, b, and c), and then aggregate these linked accounts using consolidated billing.

Consolidated billing allows you to consolidate payment for multiple member AWS accounts under a single master account, while still providing visibility for each linked account's activity. As costs and usage are aggregated in the master account, this allows you to maximize your service volume discounts, and maximize the use of your commitment discounts (Savings Plans and Reserved Instances) to achieve the highest discounts.

AWS Control Tower can quickly set up and configure multiple AWS accounts, ensuring that governance is aligned with your organization's requirements.

**Organizational Groups and Roles**: After you develop policies, you can create logical groups and roles of users within your organization. This allows you to assign permissions and control usage. Begin with high-level groupings of people, typically this aligns with organizational units and job roles (for example, systems administrator in the IT Department, or Financial controller). The groups join people that do similar tasks and need similar access. Roles define what a group must do. For example, a systems administrator in IT requires access to create all resources, but an analytics team member only needs to create analytics resources.

**Controls — Notifications:** A common first step in implementing cost controls is to setup notifications when cost or usage events occur outside of the policies. This enables you to act quickly and verify if corrective action is required, without restricting or negatively impacting workloads or new activity. After you know the workload and environment limits, you can enforce governance. In AWS, notifications are conducted with AWS Budgets, which allows you to define a monthly budget for your AWS costs, usage, and commitment discounts (Savings Plans and Reserved Instances). You can create budgets at an aggregate cost level (for example, all costs), or at a more granular level where you include only specific dimensions such as linked accounts, services, tags, or Availability Zones. You can also attach email notifications to your budgets, which will trigger when current or forecasted costs or usage exceeds a defined percentage threshold.

**Controls — Enforcement:** As a second step, you can enforce governance policies in AWS through AWS Identity and Access Management (IAM), and AWS Organizations Service Control Policies (SCP). IAM allows you to securely manage access to AWS

services and resources. Using IAM, you can control who can create and manage AWS resources, the type of resources that can be created, and where they can be created. This minimizes the creation of resources that are not required. Use the roles and groups created previously, and assign IAM policies to enforce the correct usage. SCP offers central control over the maximum available permissions for all accounts in your organization, ensuring that your accounts stay within your access control guidelines. SCPs are available only in an organization that has all features enabled, and you can configure the SCPs to either deny or allow actions for member accounts by default. Refer to the Well-Architected Security Pillar whitepaper for more details on implementing access management.

**Controls — Service Quotas:** Governance can also be implemented through management of Service Quotas. By ensuring Service Quotas are set with minimum overhead and accurately maintained, you can minimize resource creation outside of your organization's requirements. To achieve this, you must understand how quickly your requirements can change, understand projects in progress (both creation and decommission of resources) and factor in how fast quota changes can be implemented. Service Quotas can be used to increase your quotas when required.

AWS Cost Management services are integrated with the AWS Identity and Access Management (IAM) service. You use the IAM service in conjunction with Cost Management services to control access to your financial data and to the AWS tools in the billing console.

**Track workload lifecycle:** Ensure that you track the entire lifecycle of the workload. This ensures that when workloads or workload components are no longer required, they can be decommissioned or modified. This is especially useful when you release new services or features. The existing workloads and components may appear to be in use, but should be decommissioned to redirect customers to the new service. Notice previous stages of workloads — after a workload is in production, previous environments can be decommissioned or greatly reduced in capacity until they are required again.

AWS provides a number of management and governance services you can use for entity lifecycle tracking. You can use AWS Config or AWS Systems Manager to provide a detailed inventory of your AWS resources and configuration. It is recommended that you integrate with your existing project or asset management systems to keep track of active projects and products within your organization. Combining your current system with the rich set of events and metrics provided by AWS allows you to build a view of

significant lifecycle events and proactively manage resources to reduce unnecessary costs.

Refer to the Well-Architected Operational Excellence Pillar whitepaper for more details on implementing entity lifecycle tracking.

## Monitor Cost and Usage

Enable teams to take action on their cost and usage through detailed visibility into the workload. Cost optimization begins with a granular understanding of the breakdown in cost and usage, the ability to model and forecast future spend, usage, and features, and the implementation of sufficient mechanisms to align cost and usage to your organization's objectives. The following are required areas for monitoring your cost and usage:

**Configure detailed data sources:** Enable hourly granularity in Cost Explorer and create a Cost and Usage Report (CUR). These data sources provide the most accurate view of cost and usage across your entire organization. The CUR provides daily or hourly usage granularity, rates, costs, and usage attributes for all chargeable AWS services. All possible dimensions are in the CUR including: tagging, location, resource attributes, and account IDs.

Configure your CUR with the following customizations:

- Include resource IDs
- Automatically refresh the CUR
- Hourly granularity
- Versioning: Overwrite existing report
- Data integration: Athena (Parquet format and compression)

Use AWS Glue to prepare the data for analysis, and use Amazon Athena to perform data analysis, using SQL to query the data. You can also use Amazon QuickSight to build custom and complex visualizations and distribute them throughout your organization.

**Identify cost attribution categories:** Work with your finance team and other relevant stakeholders to understand the requirements of how costs must be allocated within your organization. Workload costs must be allocated throughout the entire lifecycle, including development, testing, production, and decommissioning. Understand how the costs

incurred for learning, staff development, and idea creation are attributed in the organization. This can be helpful to correctly allocate accounts used for this purpose to training and development budgets, instead of generic IT cost budgets.

**Establish workload metrics:** Understand how your workload's output is measured against business success. Each workload typically has a small set of major outputs that indicate performance. If you have a complex workload with many components, then you can prioritize the list, or define and track metrics for each component. Work with your teams to understand which metrics to use. This unit will be used to understand the efficiency of the workload, or the cost for each business output.

**Assign organization meaning to cost and usage:** Implement tagging in AWS to add organization information to your resources, which will then be added to your cost and usage information. . A tag is a key-value pair— the key is defined and must be unique across your organization, and the value is unique to a group of resources. An example of a key-value pair is the key is Environment, with a value of Production. All resources in the production environment will have this key-value pair. Tagging allows you categorize and track your costs with meaningful, relevant organization information. You can apply tags that represent organization categories (such as cost centers, application names, projects, or owners), and identify workloads and characteristics of workloads (such as, test or production) to attribute your costs and usage throughout your organization.

When you apply tags to your AWS resources (such as EC2 instances or Amazon S3 buckets) and activate the tags, AWS adds this information to your Cost and Usage Reports. You can run reports and perform analysis, on tagged and untagged resources to allow greater compliance with internal cost management policies, and ensure accurate attribution.

Creating and implementing an AWS tagging standard across your organization's accounts enables you to manage and govern your AWS environments in a consistent and uniform manner. Use Tag Policies in AWS Organizations to define rules for how tags can be used on AWS resources in your accounts in AWS Organizations. Tag Policies allow you to easily adopt a standardized approach for tagging AWS resources.

AWS Tag Editor allows you to add, delete, and manage tags of multiple resources.

AWS Cost Categories allows you to assign organization meaning to your costs, without requiring tags on resources. You can map your cost and usage information to unique internal organization structures. You define category rules to map and categorize costs using billing dimensions, such as accounts and tags. This provides another level of

management capability in addition to tagging. You can also map specific accounts and tags to multiple projects.

**Configure billing and cost optimization tools:** To modify usage and adjust costs, each person in your organization must have access to their cost and usage information. It is recommended that all workloads and teams have the following tooling configured when they use the cloud:

- **Reports:** Summarize of all cost and usage information.
- **Notifications:** Provide notifications when cost or usage is outside of defined limits.
- **Current State:** Configure a dashboard showing current levels of cost and usage. The dashboard should be available in a highly visible place within the work environment (similar to an operations dashboard).
- **Trending:** Provide the capability to show the variability in cost and usage over the required period of time, with the required granularity.
- **Forecasts:** Provide the capability to show estimated future costs.
- **Tracking:** Show the current cost and usage against configured goals or targets.
- **Analysis:** Provide the capability for team members to perform custom and deep analysis down to the hourly granularity, with all possible dimensions.

You can use AWS native tooling, such as AWS Cost Explorer, AWS Budgets, and Amazon Athena with QuickSight to provide this capability. You can also use third-party tooling, however, you must ensure that the costs of this tooling provide value to your organization.

**Allocate costs based on workload metrics:** Cost Optimization is delivering business outcomes at the lowest price point, which can only be achieved by allocating workload costs by workload metrics (measured by workload efficiency). Monitor the defined workload metrics through log files or other application monitoring. Combine this data with the workload costs, which can be obtained by looking at costs with a specific tag value or account ID. It is recommended to perform this analysis at the hourly level. Your efficiency will typically change if you have some static cost components (for example, a backend database running 24/7) with a varying request rate (for example, usage peaks at 9am – 5pm, with few requests at night). Understanding the relationship between the static and variable costs will help you to focus your optimization activities.

# Decommission Resources

After you manage a list of projects, employees, and technology resources over time you will be able to identify which resources are no longer being used, and which projects that no longer have an owner.

**Track resources over their lifetime:** Decommission workload resources that are no longer required. A common example is resources used for testing, after testing has been completed, the resources can be removed. Tracking resources with tags (and running reports on those tags) will help you identify assets for decommission. Using tags is an effective way to track resources, by labeling the resource with its function, or a known date when it can be decommissioned. Reporting can then be run on these tags. Example values for feature tagging are “featureX testing” to identify the purpose of the resource in terms of the workload lifecycle.

**Implement a decommissioning process:** Implement a standardized process across your organization to identify and remove unused resources. The process should define the frequency searches are performed, and the processes to remove the resource to ensure that all organization requirements are met.

**Decommission resources:** The frequency and effort to search for unused resources should reflect the potential savings, so an account with a small cost should be analyzed less frequently than an account with larger costs. Searches and decommission events can be triggered by state changes in the workload, such as a product going end of life or being replaced. Searches and decommission events may also be triggered by external events, such as changes in market conditions or product termination.

**Decommission resources automatically:** Use automation to reduce or remove the associated costs of the decommissioning process. Designing your workload to perform automated decommissioning will reduce the overall workload costs during its lifetime. You can use AWS Auto Scaling to perform the decommissioning process. You can also implement custom code using the API or SDK to decommission workload resources automatically.

## Resources

Refer to the following resources to learn more about AWS best practices for expenditure awareness.

- AWS Tagging Strategies
- Activating User-Defined Cost Allocation Tags

- AWS Billing and Cost Management
- Cost Management Blog
- Multiple Account Billing Strategy
- AWS SDK and Tools
- Tagging best practices
- Well-Architected Labs - Cost Fundamentals
- Well-Architected Labs – Expenditure Awareness

# Cost Effective Resources

Using the appropriate services, resources, and configurations for your workloads is key to cost savings. Consider the following when creating cost-effective resources:

- Evaluate cost when selecting services
- Select the correct resource type, size, and number
- Select the best pricing model
- Plan for data transfer

You can use AWS Solutions Architects, AWS Solutions, AWS Reference Architectures, and APN Partners to help you choose an architecture based on what you have learned.

## Evaluate Cost When Selecting Services

**Identify organization requirements:** When selecting services for your workload, it is key that you understand your organization priorities. Ensure that you have a balance between cost and other Well-Architected pillars, such as performance and reliability. A fully cost-optimized workload is the solution that is most aligned to your organization's requirements, not necessarily the lowest cost. Meet with all teams within your organization to collect information, such as product, business, technical and finance.

**Analyze all workload components:** Perform a thorough analysis on all components in your workload. Ensure that balance between the cost of analysis and the potential savings in the workload over its lifecycle. You must find the current impact, and potential future impact, of the component. For example, if the cost of the proposed resource is $10/month, and under forecasted loads would not exceed $15/month,

spending a day of effort to reduce costs by 50% ($5 a month) could exceed the potential benefit over the life of the system. Using a faster and more efficient data-based estimation will create the best overall outcome for this component.

Workloads can change over time, the right set of services may not be optimal if the workload architecture or usage changes. Analysis for selection of services must incorporate current and future workload states and usage levels. Implementing a service for future workload state or usage may reduce overall costs by reducing or removing the effort required to make future changes.

AWS Cost Explorer and the CUR can analyze the cost of a Proof of Concept (PoC) or running environment. You can also use the AWS Simple Monthly Calculator or the AWS Pricing Calculator to estimate workload costs.

**Managed Services:** Managed services remove the operational and administrative burden of maintaining a service, which allows you to focus on innovation. Additionally, because managed services operate at cloud scale, they can offer a lower cost per transaction or service.

Consider the time savings that will allow your team to focus on retiring technical debt, innovation, and value-adding features. For example, you might need to “lift and shift” your on-premises environment to the cloud as rapidly as possible and optimize later. It is worth exploring the savings you could realize by using managed services that remove or reduce license costs.

Usually, managed services have attributes that you can set to ensure sufficient capacity. You must set and monitor these attributes so that your excess capacity is kept to a minimum and performance is maximized. You can modify the attributes of AWS Managed Services using the AWS Management Console or AWS APIs and SDKs to align resource needs with changing demand. For example, you can increase or decrease the number of nodes on an Amazon EMR cluster (or an Amazon Redshift cluster) to scale out or in.

You can also pack multiple instances on an AWS resource to enable higher density usage. For example, you can provision multiple small databases on a single Amazon Relational Database Service (Amazon RDS) DB instance. As usage grows, you can migrate one of the databases to a dedicated RDS DB instance using a snapshot and restore process.

When provisioning workloads on managed services, you must understand the requirements of adjusting the service capacity. These requirements are typically time,

effort, and any impact to normal workload operation. The provisioned resource must allow time for any changes to occur, provision the required overhead to allow this. The ongoing effort required to modify services can be reduced to virtually zero by using APIs and SDKs that are integrated with system and monitoring tools, such as Amazon CloudWatch.

Amazon Relational Database Service (RDS), Amazon Redshift, and Amazon ElastiCache provide a managed database service. Amazon Athena, Amazon Elastic Map Reduce (EMR), and Amazon Elasticsearch provide a managed analytics service.

AWS Managed Services (AMS) is a service that operates AWS infrastructure on behalf of enterprise customers and partners. It provides a secure and compliant environment that you can deploy your workloads onto. AMS uses enterprise cloud operating models with automation to allow you to meet your organization requirements, move into the cloud faster, and reduce your on-going management costs.

**Serverless or Application-level Services:** You can use serverless or application-level services such as AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), and Amazon Simple Email Service (Amazon SES). These services remove the need for you to manage a resource, and provide the function of code execution, queuing services, and message delivery. The other benefit is that they scale in performance and cost in line with usage, allowing efficient cost allocation and attribution.

For more information on Serverless, refer to the Well-Architected Serverless Application lens whitepaper.

**Analyze the workload for different usage over time:** As AWS releases new services and features, the optimal services for your workload may change. Effort required should reflect potential benefits. Workload review frequency depends on your organization requirements. If it is a workload of significant cost, implementing new services sooner will maximize cost savings, so more frequent review can be advantageous. Another trigger for review is change in usage patterns. Significant changes in usage can indicate that alternate services would be more optimal. For example, for higher data transfer rates a direct connect service may be cheaper than a VPN, and provide the required connectivity. Predict the potential impact of service changes, so you can monitor for these usage level triggers and implement the most cost effective services sooner.

**Licensing costs:** The cost of software licenses can be eliminated through the use of open source software. This can have significant impact on workload costs as the size of the workload scales. Measure the benefits of licensed software against the total cost to

ensure that you have the most optimized workload. Model any changes in licensing and how they would impact your workload costs. If a vendor changes the cost of your database license, investigate how that impacts the overall efficiency of your workload. Consider historical pricing announcements from your vendors for trends of licensing changes across their products. Licensing costs may also scale independently of throughput or usage, such as licenses that scale by hardware (CPU bound licenses). These licenses should be avoided because costs can rapidly increase without corresponding outcomes.

You can use AWS License Manager to manage the software licenses in your workload. You can configure licensing rules and enforce the required conditions to help prevent licensing violations, and also reduce costs due to license overages.

## Select the Correct Resource Type, Size, and Number

By selecting the best resource type, size, and number of resources, you meet the technical requirements with the lowest cost resource. Right-sizing activities takes into account all of the resources of a workload, all of the attributes of each individual resource, and the effort involved in the right-sizing operation. Right-sizing can be an iterative process, triggered by changes in usage patterns and external factors, such as AWS price drops or new AWS resource types. Right-sizing can also be one-off if the cost of the effort to right-size, out weights the potential savings over the life of the workload.

In AWS, there are a number of different approaches:

- Perform cost modeling
- Select size based on metrics or data
- Select size automatically (based on metrics)

**Cost Modeling:** Perform cost modeling for your workload and each of its components to understand the balance between resources, and find the correct size for each resource in the workload, given a specific level of performance. Perform benchmark activities for the workload under different predicted loads and compare the costs. The modeling effort should reflect potential benefit; for example, time spent is proportional to component cost or predicted saving. For best practices, refer to the Review section of the Performance Efficiency Pillar of the AWS Well-Architected Framework whitepaper.

AWS Compute Optimizer can assist with cost modeling for running workloads. It provides right-sizing recommendations for compute resources based on historical

usage. This is the ideal data source for compute resources because it is a free service, and it utilizes machine learning to make multiple recommendations depending on levels of risk. You can also use Amazon CloudWatch and CloudWatch Logs with custom logs as data sources for right sizing operations for other services and workload components.

The following are recommendations for cost modeling data and metrics:

- The monitoring must accurately reflect the end-user experience. Select the correct granularity for the time period and thoughtfully choose the maximum or 99th percentile instead of the average.
- Select the correct granularity for the time period of analysis that is required to cover any workload cycles. For example, if a two-week analysis is performed, you might be overlooking a monthly cycle of high utilization, which could lead to under-provisioning.

**Metrics or data-based selection:** Select resource size or type based on workload and resource characteristics; for example, compute, memory, throughput, or write intensive. This selection is typically made using cost modeling, a previous version of the workload (such as an on-premises version), using documentation, or using other sources of information about the workload (whitepapers, published solutions).

**Automatic selection based on metrics:** Create a feedback loop within the workload that uses active metrics from the running workload to make changes to that workload. You can use a managed service, such as AWS Auto Scaling, which you configure to perform the right sizing operations for you. AWS also provides APIs, SDKs, and features that allow resources to be modified with minimal effort. You can program a workload to stop-and-start an EC2 instance to allow a change of instance size or instance type. This provides the benefits of right-sizing while removing almost all the operational cost required to make the change.

Some AWS services have built in automatic type or size selection, such as S3 Intelligent-Tiering. S3 Intelligent-Tiering automatically moves your data between two access tiers: frequent access and infrequent access, based on your usage patterns.

## Select the Best Pricing Model

**Perform workload cost modeling:** Consider the requirements of the workload components and understand the potential pricing models. Define the availability requirement of the component. Determine if there are multiple independent resources that perform the function in the workload, and what the workload requirements are over time. Compare the cost of the resources using the default On-Demand pricing model

and other applicable models. Factor in any potential changes in resources or workload components.

**Perform regular account level analysis:** Performing regular cost modeling ensures that opportunities to optimize across multiple workloads can be implemented. For example, if multiple workloads use On-Demand, at an aggregate level, the risk of change is lower, and implementing a commitment-based discount will achieve a lower overall cost. It is recommended to perform analysis in regular cycles of two weeks to 1 month. This allows you to make small adjustment purchases, so the coverage of your pricing models continues to evolve with your changing workloads and their components.

Use the AWS Cost Explorer recommendations tool to find opportunities for commitment discounts.

To find opportunities for Spot workloads, use an hourly view of your overall usage, and look for regular periods of changing usage or elasticity.

**Pricing Models:** AWS has multiple pricing models that allow you to pay for your resources in the most cost-effective way that suits your organization's needs. The following section describes each purchasing model:

- On-Demand
- Spot
- Commitment discounts - Savings Plans
- Commitment discounts - Reserved Instances/Capacity
- Geographic selection
- Third-party agreements and pricing

**On-Demand:** This is the default, pay as you go pricing model. When you use resources (for example, EC2 instances or services such as DynamoDB on demand) you pay a flat rate, and you have no long-term commitments. You can increase or decrease the capacity of your resources or services based on the demands of your application. On-Demand has an hourly rate, but depending on the service, can be billed in increments of 1 second (for example Amazon RDS, or Linux EC2 instances). On demand is recommended for applications with short-term workloads (for example, a four-month project), that spike periodically, or unpredictable workloads that can't be interrupted. On demand is also suitable for workloads, such as pre-production environments, which require uninterrupted runtimes, but do not run long enough for a commitment discount (Savings Plans or Reserved Instances).

**Spot:** A Spot Instance is spare EC2 compute capacity available at discounts of up to 90% off On-Demand prices with no long-term commitment required. With Spot Instances, you can significantly reduce the cost of running your applications or scale your application's compute capacity for the same budget. Unlike On-Demand, Spot Instances can be interrupted with a 2-minute warning if EC2 needs the capacity back, or the Spot Instance price exceeds your configured price. On average, Spot Instances are interrupted less than 5% of the time.

Spot is ideal when there is a queue or buffer in place, or where there are multiple resources working independently to process the requests (for example, Hadoop data processing). Typically these workloads are fault-tolerant, stateless, and flexible, such as batch processing, big data and analytics, containerized environments, and high performance computing (HPC). Non-critical workloads such as test and development environments are also candidates for Spot.

Spot is also integrated into multiple AWS services, such as EC2 Auto Scaling groups (ASGs), Elastic MapReduce (EMR), Elastic Container Service (ECS), and AWS Batch.

When a Spot Instance needs to be reclaimed, EC2 sends a two-minute warning via a Spot Instance interruption notice delivered through CloudWatch Events, as well as in the instance metadata. During that two-minute period, your application can use the time to save its state, drain running containers, upload final log files, or remove itself from a load balancer. At the end of the two minutes, you have the option to hibernate, stop, or terminate the Spot Instance.

Consider the following best practices when adopting Spot Instances in your workloads:

- **Set your maximum price as the On-Demand rate:** This ensures that you will pay the current spot rate (the cheapest available price) and will never pay more than the On-Demand rate. Current and historical rates are available via the console and API.
- **Be flexible across as many instance types as possible:** Be flexible in both the family and size of the instance type, to improve the likelihood of fulfilling your target capacity requirements, obtain the lowest possible cost, and minimize the impact of interruptions.
- **Be flexible about where your workload will run:** Available capacity can vary by Availability Zone. This improves the likelihood of fulfilling your target capacity by tapping into multiple spare capacity pools, and provides the lowest possible cost.

- **Design for continuity:** Design your workloads for statelessness and fault-tolerance, so that if some of your EC2 capacity gets interrupted, it will not have impact on the availability or performance of the workload.
- We recommend using Spot Instances in combination with On-Demand and Savings Plans/Reserved Instances to maximize workload cost optimization with performance.

**Commitment Discounts – Savings Plans:** AWS provides a number of ways for you to reduce your costs by reserving or committing to use a certain amount of resources, and receiving a discounted rate for your resources. A Savings Plan allows you to make an hourly spend commitment for one or three years, and receive discounted pricing across your resources. Savings Plans provide discounts for AWS Compute services such as EC2, Fargate, and Lambda. When you make the commitment, you pay that commitment amount every hour, and it is subtracted from your On-Demand usage at the discount rate. For example, you commit to $50 an hour, and have $150 an hour of On-Demand usage. Considering the Savings Plans pricing, your specific usage has a discount rate of 50%. So, your $50 commitment covers $100 of On-Demand usage. You will pay $50 (commitment) and $50 of remaining On-Demand usage.

Compute Savings Plans are the most flexible and provide a discount of up to 66%. They automatically apply across Availability Zones, instance size, instance family, operating system, tenancy, Region, and compute service.

Instance Savings Plans have less flexibility but provide a higher discount rate (up to 72%). They automatically apply across Availability Zones, instance size, instance family, operating system, and tenancy.

There are three payment options:

- **No upfront payment:** There is no upfront payment; you then pay a reduced hourly rate each month for the total hours in the month.
- **Partial upfront payment:** Provides a higher discount rate than No upfront. Part of the usage is paid up front; you then pay a smaller reduced hourly rate each month for the total hours in the month.
- **All upfront payment:** Usage for the entire period is paid up front, and no other costs are incurred for the remainder of the term for usage that is covered by the commitment.

You can apply any combination of these three purchasing options across your workloads.

Savings plans apply first to the usage in the account they are purchased in, from the highest discount percentage to the lowest, then they apply to the consolidated usage across all other accounts, from the highest discount percentage to the lowest.

It is recommended to purchase all Savings Plans in an account with no usage or resources, such as the master account. This ensures that the Savings Plan applies to the highest discount rates across all of your usage, maximizing the discount amount.

Workloads and usage typically change over time. It is recommended to continually purchase small amounts of Savings Plans commitment over time. This ensures that you maintain high levels of coverage to maximize your discounts, and your plans closely match your workload and organization requirements at all times.

Do not set a target coverage in your accounts, due to the variability of discount that is possible. Low coverage does not necessarily indicate high potential savings. You may have a low coverage in your account, but if your usage is made up of small instances, with a licensed operating system, the potential saving could be as low as a few percent. Instead, track and monitor the potential savings available in the Savings Plan recommendation tool. Frequently review the Savings Plans recommendations in Cost Explorer (perform regular analysis) and continue to purchase commitments until the estimated savings are below the required discount for the organization. For example, track and monitor that your potential discounts remained below 20%, if it goes above that a purchase must be made.

Monitor the utilization and coverage, but only to detect changes. Do not aim for a specific utilization percent, or coverage percent, as this does not necessarily scale with savings. Ensure that a purchase of Savings Plans results in an increase in coverage, and if there are decreases in coverage or utilization ensure they are quantified and known. For example, you migrate a workload resource to a newer instance type, which reduces utilization of an existing plan, but the performance benefit outweighs the saving reduction.

**Commitment Discounts – Reserved Instances/Commitment:** Similar to Savings Plans, Reserved Instances (RI) offer discounts up to 72% for a commitment to running a minimum amount of resources. Reserved Instances are available for RDS, Elasticsearch, ElastiCache, Amazon Redshift, and DynamoDB. Amazon CloudFront and AWS Elemental MediaConvert also provide discounts when you make minimum usage commitments. Reserved Instances are currently available for EC2, however Savings Plans offer the same discount levels with increased flexibility and no management overhead.

Reserved Instances offer the same pricing options of no upfront, partial upfront, and all upfront, and the same terms of one or three years.

Reserved Instances can be purchased in a Region or a specific Availability Zone. They provide a capacity reservation when purchased in an Availability Zone,.

EC2 features convertible RI's, however, Savings Plans should be used for all EC2 instances due to increased flexibility and reduced operational costs.

The same process and metrics should be used to track and make purchases of Reserved Instances. It is recommended to not track coverage of RI's across your accounts. It is also recommended that utilization % is not monitored or tracked, instead view the utilization report in Cost Explorer, and use net savings column in the table. If the net savings is a significantly large negative amount, you must take action to remediate the unused RI.

**EC2 Fleet:** EC2 Fleet is a feature that allows you to define a target compute capacity, and then specify the instance types and the balance of On-Demand and Spot for the fleet. EC2 Fleet will automatically launch the lowest price combination of resources to meet the defined capacity.

**Geographic Selection:** When you architect your solutions, a best practice is to seek to place computing resources closer to users to provide lower latency and strong data sovereignty. For global audiences, you should use multiple locations to meet these needs. You should select the geographic location that minimizes your costs.

The AWS Cloud infrastructure is built around Regions and Availability Zones. A Region is a physical location in the world where we have multiple Availability Zones. Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities.

Each AWS Region operates within local market conditions, and resource pricing is different in each Region. Choose a specific Region to operate a component of or your entire solution so that you can run at the lowest possible price globally. You can use the AWS Simple Monthly Calculator to estimate the costs of your workload in various Regions.

**Third-party agreements and pricing:** When you utilize third-party solutions or services in the cloud, it is important that the pricing structures are aligned to Cost Optimization outcomes. Pricing should scale with the outcomes and value it provides. An example of this is software that takes a percentage of savings it provides, the more you save (outcome) the more it charges. Agreements that scale with your bill are typically not

aligned to Cost Optimization, unless they provide outcomes for every part of your specific bill. For example, a solution that provides recommendations for EC2 and charges a percentage of your entire bill will increase if you use other services for which it provides no benefit. Another example is a managed service that is charged at a percentage of the cost of resources that are managed. A larger instance size may not necessarily require more management effort, but will be charged more. Ensure that these service pricing arrangements include a cost optimization program or features in their service to drive efficiency.

## Plan for Data Transfer

An advantage of the cloud is that it is a managed network service. There is no longer the need to manage and operate a fleet of switches, routers, and other associated network equipment. Networking resources in the cloud are consumed and paid for in the same way you pay for CPU and storage—you only pay for what you use. Efficient use of networking resources is required for cost optimization in the cloud.

**Perform data transfer modeling:** Understand where the data transfer occurs in your workload, the cost of the transfer, and its associated benefit. This allows you to make an informed decision to modify or accept the architectural decision. For example, you may have a Multi-Availability Zone configuration where you replicate data between the Availability Zones. You model the cost of structure and decide that this is an acceptable cost (similar to paying for compute and storage in both Availability Zone) to achieve the required reliability and resilience.

Model the costs over different usage levels. Workload usage can change over time, and different services may be more cost effective at different levels.

Use AWS Cost Explorer or the Cost and Usage Report (CUR) to understand and model your data transfer costs. Configure a proof of concept (PoC) or test your workload, and run a test with a realistic simulated load. You can model your costs at different workload demands.

**Optimize Data Transfer:** Architecting for data transfer ensures that you minimize data transfer costs. This may involve using content delivery networks to locate data closer to users, or using dedicated network links from your premises to AWS. You can also use WAN optimization and application optimization to reduce the amount of data that is transferred between components.

**Select services to reduce data transfer costs:** Amazon CloudFront is a global content delivery network that delivers data with low latency and high transfer speeds. It

caches data at edge locations across the world, which reduces the load on your resources. By using CloudFront, you can reduce the administrative effort in delivering content to large numbers of users globally, with minimum latency.

AWS Direct Connect allows you to establish a dedicated network connection to AWS. This can reduce network costs, increase bandwidth, and provide a more consistent network experience than internet-based connections.

AWS VPN allows you to establish a secure and private connection between your private network and the AWS global network. It is ideal for small offices or business partners because it provides quick and easy connectivity, and it is a fully managed and elastic service.

VPC Endpoints allow connectivity between AWS services over private networking and can be used to reduce public data transfer and NAT gateways costs. Gateway VPC endpoints have no hourly charges, and support Amazon S3 and Amazon DynamoDB. Interface VPC endpoints are provided by AWS PrivateLink and have an hourly fee and per GB usage cost.

## Resources

Refer to the following resources to learn more about AWS best practices for cost effective resources.

- AWS Managed Services: Enterprise Transformation Journey Video
- Analyzing Your Costs with Cost Explorer
- Accessing Reserved Instance Recommendations
- Getting Started with Rightsizing Recommendations
- Spot Instances Best Practices
- Spot Fleets
- How Reserved Instances Work
- AWS Global Infrastructure
- Spot Instance Advisor
- Well-Architected Labs - Cost Effective Resources

# Manage Demand and Supply Resources

When you move to the cloud, you pay only for what you need. You can supply resources to match the workload demand at the time they're needed — eliminating the need for costly and wasteful overprovisioning. You can also modify the demand using a throttle, buffer, or queue to smooth the demand and serve it with less resources.

The economic benefits of just-in-time supply should be balanced against the need to provision to account for resource failures, high availability, and provision time. Depending on whether your demand is fixed or variable, plan to create metrics and automation that will ensure that management of your environment is minimal – even as you scale. When modifying the demand, you must know the acceptable and maximum delay that the workload can allow.

In AWS, you can use a number of different approaches for managing demand and supplying resources. The following sections describe how to use these approaches:

- Analyze the workload
- Manage demand
- Demand-based supply
- Time-based supply

**Analyze the workload:** Know the requirements of the workload. The organization requirements should indicate the workload response times for requests. The response time can be used to determine if the demand is managed, or if the supply of resources will change to meet the demand.

The analysis should include the predictability and repeatability of the demand, the rate of change in demand, and the amount of change in demand. Ensure that the analysis is performed over a long enough period to incorporate any seasonal variance, such as end-of-month processing or holiday peaks.

Ensure that the analysis effort reflects the potential benefits of implementing scaling. Look at the expected total cost of the component, and any increases or decreases in usage and cost over the workload lifetime.

You can use AWS Cost Explorer or Amazon QuickSight with the CUR or your application logs to perform a visual analysis of workload demand.

## Manage Demand

**Manage Demand – Throttling:** If the source of the demand has retry capability, then you can implement throttling. Throttling tells the source that if it cannot service the request at the current time it should try again later. The source will wait for a period of time and then re-try the request. Implementing throttling has the advantage of limiting the maximum amount of resources and costs of the workload. In AWS, you can use Amazon API Gateway to implement throttling. Refer to the Well-Architected Reliability pillar whitepaper for more details on implementing throttling.

**Manage Demand – Buffer based:** Similar to throttling, a buffer defers request processing, allowing applications that run at different rates to communicate effectively. A buffer-based approach uses a queue to accept messages (units of work) from producers. Messages are read by consumers and processed, allowing the messages to run at the rate that meets the consumers' business requirements. You don't have to worry about producers having to deal with throttling issues, such as data durability and backpressure (where producers slow down because their consumer is running slowly).

In AWS, you can choose from multiple services to implement a buffering approach. Amazon SQS is a managed service that provides queues that allow a single consumer to read individual messages. Amazon Kinesis provides a stream that allows many consumers to read the same messages.

When architecting with a buffer-based approach, ensure that you architect your workload to service the request in the required time, and that you are able to handle duplicate requests for work.

## Dynamic Supply

**Demand-based supply:** Leverage the elasticity of the cloud to supply resources to meet changing demand. Take advantage of APIs or service features to programmatically vary the amount of cloud resources in your architecture dynamically. This allows you to scale components in your architecture, and automatically increase the number of resources during demand spikes to maintain performance, and decrease capacity when demand subsides to reduce costs.

Auto Scaling helps you adjust your capacity to maintain steady, predictable performance at the lowest possible cost. It is a fully managed and free service that integrates with Amazon EC2 instances and Spot Fleets, Amazon ECS, Amazon DynamoDB, and Amazon Aurora.

Auto Scaling provides automatic resource discovery to help find resources in your workload that can be configured, it has built-in scaling strategies to optimize performance, costs or a balance between the two, and provides predictive scaling to assist with regularly occurring spikes.

Auto Scaling can implement manual, scheduled or demand based scaling, you can also use metrics and alarms from Amazon CloudWatch to trigger scaling events for your workload. Typical metrics can be standard Amazon EC2 metrics, such as CPU utilization, network throughput, and ELB observed request/response latency. When possible, you should use a metric that is indicative of customer experience, typically this a custom metric that might originate from application code within your workload.

When architecting with a demand-based approach keep in mind two key considerations. First, understand how quickly you must provision new resources. Second, understand that the size of margin between supply and demand will shift. You must be ready to cope with the rate of change in demand and also be ready for resource failures.

Elastic Load Balancing (ELB) helps you to scale by distributing demand across multiple resources. As you implement more resources, you add them to the load balancer to take on the demand. AWS ELB has support for EC2 Instances, containers, IP addresses ,and Lambda functions.

**Time-based supply:** A time-based approach aligns resource capacity to demand that is predictable or well-defined by time. This approach is typically not dependent upon utilization levels of the resources. A time-based approach ensures that resources are available at the specific time they are required, and can be provided without any delays due to start-up procedures and system or consistency checks. Using a time-based approach, you can provide additional resources or increase capacity during busy periods.

You can use scheduled Auto Scaling to implement a time-based approach. Workloads can be scheduled to scale out or in at defined times (for example, the start of business hours) thus ensuring that resources are available when users or demand arrives.

You can also leverage the AWS APIs and SDKs and AWS CloudFormation to automatically provision and decommission entire environments as you need them. This approach is well suited for development or test environments that run only in defined business hours or periods of time.

You can use APIs to scale the size of resources within an environment (vertical scaling). For example, you could scale up a production workload by changing the instance size

or class. This can be achieved by stopping and starting the instance and selecting the different instance size or class. This technique can also be applied to other resources, such as EBS Elastic Volumes, which can be modified to increase size, adjust performance (IOPS) or change the volume type while in use.

When architecting with a time-based approach keep in mind two key considerations. First, how consistent is the usage pattern? Second, what is the impact if the pattern changes? You can increase the accuracy of predictions by monitoring your workloads and by using business intelligence. If you see significant changes in the usage pattern, you can adjust the times to ensure that coverage is provided.

**Dynamic Supply:** You can use AWS Auto Scaling, or incorporate scaling in your code with the AWS API or SDKs. This reduces your overall workload costs by removing the operational cost from manually making changes to your environment, and can be performed much faster. This will ensure that the workload resourcing best matches the demand at any time.

### Resources

Refer to the following resources to learn more about AWS best practices for managing demand and supplying resources.

- API Gateway Throttling
- Getting Started with Amazon SQS
- Getting Started with Amazon EC2 Auto Scaling

# Optimize Over Time

In AWS, you optimize over time by reviewing new services and implementing them in your workload.

## Review and Implement New Services

As AWS releases new services and features, it is a best practice to review your existing architectural decisions to ensure that they remain cost effective. As your requirements change, be aggressive in decommissioning resources, components, and workloads that you no longer require. Consider the following to help you optimize over time:

- Develop a workload review process

- Review and implement services

**Develop a workload review process:** To ensure that you always have the most cost efficient workload, you must regularly review the workload to know if there are opportunities to implement new services, features, and components. To ensure that you achieve overall lower costs the process must be proportional to the potential amount of savings. For example, workloads that are 50% of your overall spend should be reviewed more regularly, and more thoroughly, than workloads that are 5% of your overall spend. Factor in any external factors or volatility. If the workload services a specific geography or market segment, and change in that area is predicted, more frequent reviews could lead to cost savings. Another factor in review is the effort to implement changes. If there are significant costs in testing and validating changes, reviews should be less frequent.

Factor in the long-term cost of maintaining outdated and legacy, components and resources, and the inability to implement new features into them. The current cost of testing and validation may exceed the proposed benefit. However, over time, the cost of making the change may significantly increase as the gap between the workload and the current technologies increases, resulting in even larger costs. For example, the cost of moving to a new programming language may not currently be cost effective. However, in five years time, the cost of people skilled in that language may increase, and due to workload growth, you would be moving an even larger system to the new language, requiring even more effort than previously.

Break down your workload into components, assign the cost of the component (an estimate is sufficient), and then list the factors (for example, effort and external markets) next to each component. Use these indicators to determine a review frequency for each workload. For example, you may have webservers as a high cost, low change effort, and high external factors, resulting in high frequency of review. A central database may be medium cost, high change effort, and low external factors, resulting in a medium frequency of review.

**Review the workload and implement services:** To realize the benefits of new AWS services and features, you must execute the review process on your workloads and implement new services and features as required. For example, you might review your workloads and replace the messaging component with Amazon Simple Email Service (SES). This removes the cost of operating and maintaining a fleet of instances, while providing all the functionality at a reduced cost.

# Conclusion

Cost optimization and Cloud Financial Management is an ongoing effort. You should regularly work with your finance and technology teams, review your architectural approach, and update your component selection.

AWS strives to help you minimize cost while you build highly resilient, responsive, and adaptive deployments. To truly optimize the cost of your deployment, take advantage of the tools, techniques, and best practices discussed in this paper.

# Contributors

Contributors to this document include:

- Philip Fitzsimons, Sr Manager Well-Architected, Amazon Web Services
- Nathan Besh, Cost Lead Well-Architected, Amazon Web Services
- Levon Stepanian, BDM, Cloud Financial Management, Amazon Web Services
- Keith Jarrett, Business Development Lead – Cost Optimization
- PT Ng, Commercial Architect, Amazon Web Services
- Arthur Basbaum, Business Developer Manager, Amazon Web Service
- Jarman Hauser, Commercial Architect, Amazon Web Services

# Further Reading

For additional information, see:

- AWS Well-Architected Framework

# Document Revisions

| Date | Description |
|---|---|
| April 2020 | Updated to incorporate CFM, new services, and integration with the Well-Architected too. |
| July 2018 | Updated to reflect changes to AWS and incorporate learnings from reviews with customers. |
| November 2017 | Updated to reflect changes to AWS and incorporate learnings from reviews with customers. |
| November 2016 | First publication |