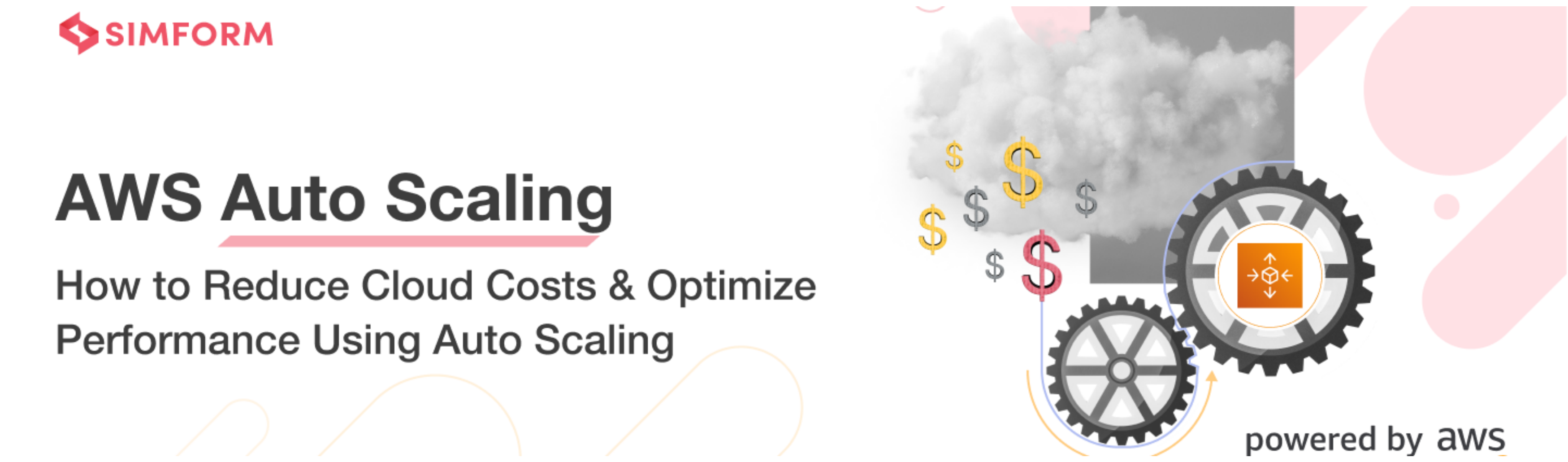
AWS Auto Scaling: Reduce cloud costs & optimize performance

Check out this article to learn how to make the most of AWS Auto Scaling and its benefits explained with real-life examples!

Hiren Dhaduk
August 17, 2022

7 mins read Last Updated September 28, 2022


SIMFORM
AWS Auto Scaling
How to Reduce Cloud Costs & Optimize Performance Using Auto Scaling
powered by aws

Quick Summary :– *If your application experiences traffic fluctuations regularly or even an unprecedented spike, how can you ensure you have enough instances to support it without burning a hole in your pocket? This article breaks down everything about AWS Auto Scaling and how it can help optimize costs and performance.*

Table of Contents

Let's say you have an e-commerce web application in AWS that experiences variable traffic throughout the day. It has a standard 3-tier architecture consisting of:

- Amazon EC2 for compute layer
- DynamoDB for the data layer
- Elastic Load Balancing for distributing incoming traffic

This is where AWS Auto Scaling steps in!

It lets you configure and manage application-wide scaling with a single, unified interface. You can set up scaling for multiple resources in your application in just a few clicks, and AWS Auto Scaling automatically adjusts the capacity.

Thus, **it maintains steady performance while driving your cloud costs** down as you only pay for the resources used.

This comprehensive guide also talks about how AWS Auto Scaling works, its use cases, and best practices for making the most out of it. Read on to know more about AWS Auto Scaling and its benefits with examples.

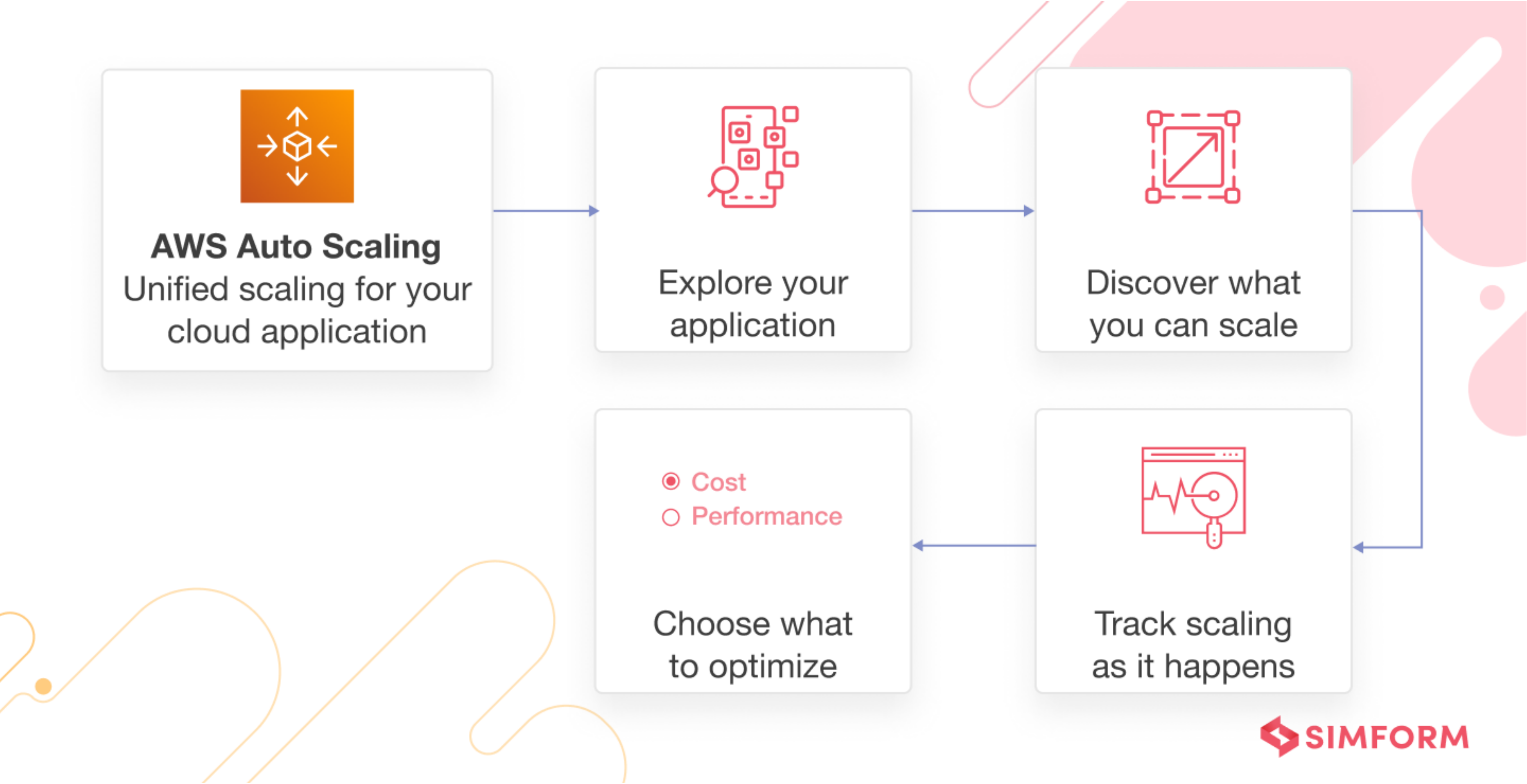
AWS Auto Scaling: An overview

With **AWS Auto Scaling** you can set up scaling for multiple resources across multiple services in your application in minutes. You no longer need to set alarms and scaling actions for each resource or service. It provides a single unified interface to configure and manage automatic scaling for all scalable resources in your application, including:

- Amazon EC2 instances
- EC2 Spot Fleets
- ECS tasks
- DynamoDB tables and Global Secondary Indexes
- Aurora Replicas

AWS Auto Scaling accomplishes this by monitoring your applications and automatically adjusting capacity as needed to align with your scaling strategy. This allows you to maintain optimum performance at the lowest possible costs.

How it works


AWS Auto Scaling
Unified scaling for your cloud application
Explore your application
Discover what you can scale
Track scaling as it happens
Cost
Performance
Choose what to optimize
SIMFORM

To sum up, the **features of AWS Auto Scaling include:**

- Simple yet powerful interface

Built-in scaling strategies designed to optimize performance, costs, or a balance of both

Predictive scaling

Typically, resources are strictly defined and constrained within set configurations for a given set of resources. For instance, if an organization has a large analytics workload it needs to process, it may require more compute and memory resources than initially defined. But with autoscaling, the resources can scale up to process data quickly.

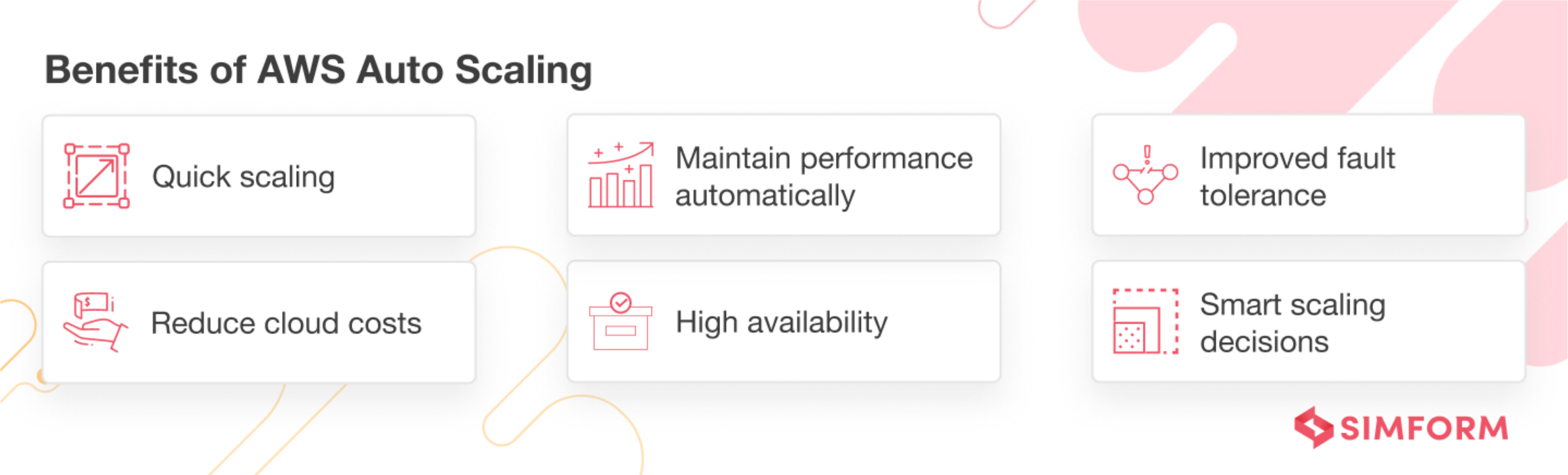
Simform engineered a UX-driven, cloud-native solution for ONA, a dating cum coaching platform. One major challenge we faced was building a flexible architecture for better scalability and app performance at scale. So, **we used AWS Auto Scaling to handle spikes in user traffic efficiently**. And we also used Elastic load balancer to seamlessly distribute incoming traffic among multiple EC2 instances to avoid server downtimes.

Want to read our full case study on ONA app development?

CLICK HERE

Benefits of AWS Auto Scaling

AWS Auto Scaling does more than optimize performance and cost for your applications. It has numerous benefits that are discussed below with real-life examples.



Quick scaling

It simplifies the scaling exercise for your teams by allowing them to scale a set of related resources with just a few clicks. In addition, it lets you see the average utilization of all your scalable resources and set target utilization levels for them from one intuitive interface.

For CoinJar, a cryptocurrency platform, scaling for maintaining transaction speed was of utmost significance given the unpredictable nature of digital currency markets. Spikes in traffic during market volatility resulted in thousands of requests per second, creating major technical challenges for its small teams. But with AWS Auto Scaling, **CoinJar could scale effectively with a team of only seven IT and DevOps engineers**.

optimize utilization and achieve cost-efficiency as you only pay for the resources your applications use. For instance, it automatically removes excess capacity when demand drops, so you do not overspend.

Conflux Technologies built a SaaS banking platform in AWS named Finflux, which experienced transactions that peaked around 10 am on weekdays and again in the evenings. But it could scale up and down as needed with AWS Auto Scaling, maintaining performance and optimizing costs. As a result, **Conflux reduced Finflux's IT spending by 40%.**

Maintain performance automatically

AWS Auto Scaling continuously monitors the resources your applications use to ensure they operate at your desired performance levels. For instance, it automatically increases the capacity of constrained resources when traffic spikes, so you maintain a high quality of service.

A great example of how AWS Auto Scaling maintains steady performance is when Pacific Magazines withstood an unprecedented **spike of 500% during a DDoS attack**. The company used AWS Auto Scaling to monitor and respond to changing capacity needs. And during the spike, Auto Scaling seamlessly scaled up until the attack was mitigated, and the users experienced zero downtime. As a result, the company suffered no revenue losses or reputational damage.

High availability

Cloud resources can become unavailable if overwhelmed with resource-intensive workloads or more traffic than the instances are set up to handle. For example, in events such as Black Friday, online app stores may experience a significant change from a typical use pattern for a service site. Here, Auto Scaling can deliver the necessary resources for the service to continue its operations and maintain availability.

MCredit, a FinTech firm, faced frequent downtimes and crashes during peak loads before migrating to AWS. Post-migration, it implemented AWS Auto Scaling to adjust capacity automatically for consistent system performance. As a result, MCredit reported higher availability and stability in AWS, allowing it to serve customers any time of the day.

Improved fault tolerance

AWS Auto Scaling constantly monitors the performance of your workloads and their health. Thus, it can terminate and replace corrupted instances by automatically launching new ones, improving reliability and fault tolerance.

Smart scaling decisions

AWS Auto Scaling automatically creates scaling policies about how groups of different resources should respond to changes in demand and sets targets based on your preferences. Moreover, it continuously calculates appropriate scaling adjustments to immediately add or remove capacity to keep your metrics on target. In addition, the target tracking policies are self-optimizing as they learn your load patterns to minimize fluctuations in resource capacity. Thus, it results in a smoother and smarter scaling experience for you.

Types of AWS Auto Scaling

AWS Auto Scaling primarily uses two types of techniques to automatically scale resources, as discussed below.

Dynamic scaling

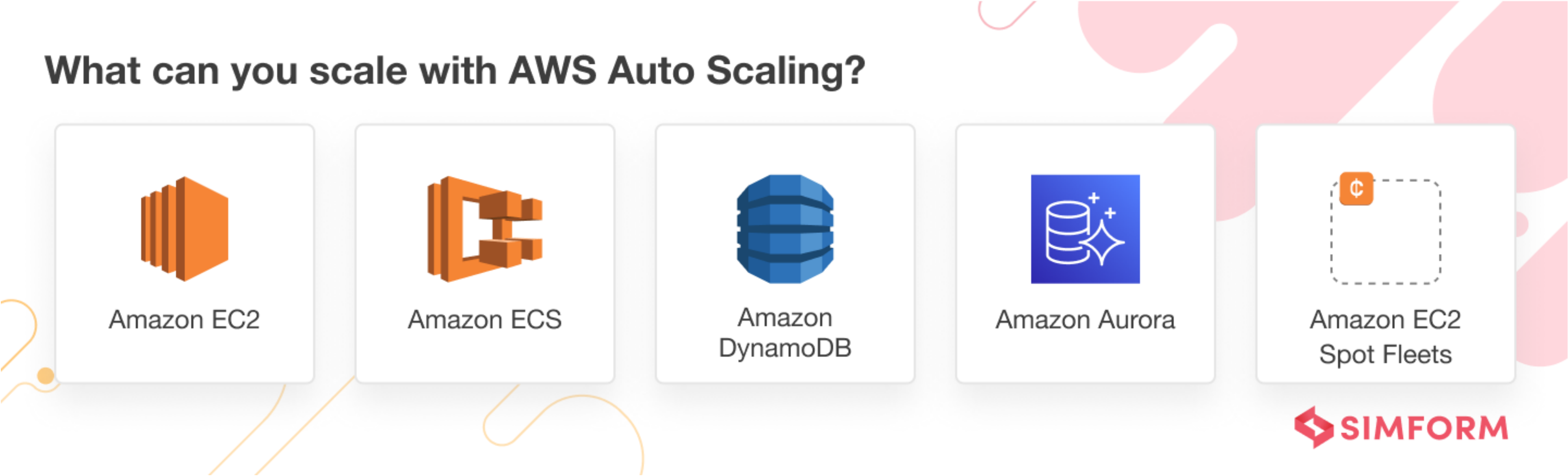
It scales resources per the traffic changes that occur and is also tightly coupled with real-time monitoring of resources.

Predictive scaling

in lower costs and more responsive applications.

What can you scale with AWS Auto Scaling?

AWS Auto Scaling does not support all AWS resources. It can be used to set up scaling for resources as discussed below.



Amazon EC2

You can launch or terminate Amazon EC2 instances in EC2 Auto Scaling groups with dynamic scaling. In addition, EC2 Auto Scaling further supports the following types of dynamic scaling policies:

> ***Target tracking scaling*** *increases or decreases current capacity based on a CloudWatch metric and a target value.*
>
> ***Step scaling*** *increases or decreases current capacity based on step adjustments (a set of scaling adjustments) that vary depending on the size of an alarm breach.*
>
> ***Simple scaling*** *increases or decreases current capacity based on a single scaling adjustment, with a cooldown period between each scaling activity.*

You may prefer **Amazon EC2 Auto Scaling** if you only need to scale EC2 groups or ensure their health. And if you are already using EC2 Auto Scaling, you can combine it with AWS Auto Scaling to scale other resources.

Kidoodle.TV, a children's streaming service, experienced a spike in usage after 2020, brought on by the COVID-19 pandemic. But the company could seamlessly scale up to meet the increase in viewership and ad-related workloads. By using AWS Auto Scaling, it could simply ramp up its use of EC2 and ensure it had the compute capacity required for its increased workloads.

Amazon EC2 Spot Fleets

An EC2 **Spot Fleet** is a set of Spot Instances mixed with optionally On-Demand Instances, launched based on criteria you specify. You can launch or terminate instances from a Spot Fleet or automatically replace instances that get interrupted for price or capacity reasons.

It uses automatic scaling, which is the ability to automatically increase and decrease the target capacity of Spot Fleets based on demand. It further supports the following types of scaling:

> ***Target tracking scaling***
>
> ***Step scaling***

Scheduled scaling increases or decreases the current capacity of a fleet based on a set date and time.

tasks in an Amazon ECS service. In addition, ECS publishes CloudWatch metrics with your services' average memory and CPU usage. You can use these metrics to add or remove tasks to deal with high demand at peak times or **reduce ECS costs** during periods of low utilization.

Amazon ECS Service Auto Scaling further supports the following types of automatic scaling:

- ***Target tracking scaling policies***
- ***Step scaling policies***
- ***Scheduled scaling***

Ula, a B2B e-commerce marketplace app for MSMEs in Indonesia, uses ECS for automated **container orchestration** at scale with ECS Spot Instances for controlling costs. It uses AWS Auto Scaling to scale resources on demand and deploys EC2 Reserved Instances for backup when Spot Instances are terminated. As a result, it maintains 99% or higher uptime, saves on monthly infrastructure costs, and can scale to support large growth in business volumes.

Amazon DynamoDB

Database workloads are either cyclical or can be challenging to predict in advance. For example, for a social networking app where most users are active during the daytime, there is no need for the same levels of throughput at night. As for a new gaming app that suddenly gains popularity, it could exceed the available database resources and result in slow performance.

DynamoDB autoscaling dynamically adjusts throughput capacity in response to actual traffic patterns. It enables a DynamoDB table or a global secondary index (to increase provisioned read and write capacity) to handle sudden traffic increases without throttling. When the traffic subsides, it decreases throughput, so you do not pay for unused provisioned capacity.

To further achieve consistent, low-latency performance for your applications while lowering infrastructure costs, read our blog post on **how to choose the right database on AWS.**

Amazon Aurora

You can use Aurora Auto Scaling to dynamically add or remove Aurora Replicas to handle changes in active connections and workloads automatically. It enables your Aurora DB clusters to handle sudden increases in connectivity or workload. When the connectivity or workload decreases, it removes unnecessary Aurora Replicas, so you do not pay for unused provisioned DB instances.

When to use AWS Auto Scaling?

If you are wondering when to use AWS Auto Scaling, below are some common use cases:

- *You have an application using one or more scalable resources that experience variable loads*
- *You need to manage scaling for multiple resources across multiple services*
- *You want to provide consistent performance for users during high-demand periods and save costs during low demand.*
- *You have applications that have periodic or unpredictable traffic variations, as predictive scaling can thrive here*
- *You have workloads with traffic flow changing on a weekly or daily basis*
- *You want to set up predictive scaling for EC2 resources*

campaigns

If you are looking to build a highly scalable app on AWS, check out our step-by-step guide on **building a scalable application for up to 1 million users on AWS**.

Best practices for AWS Auto scaling

To help you make the most with AWS Auto Scaling, below are a few tips to keep in mind when using it.

Create a custom strategy as per metrics and thresholds you define to suit your specific needs instead of the existing scaling recommendations.

If you are unsure whether you want to optimize for cost or performance, consider going with the recommended strategies.

Use EC2 Auto Scaling if your infrastructure only uses EC2, and you need to scale EC2 Auto Scaling groups.

Enhance the benefits of Auto Scaling further by using Elastic Load Balancing.

Combine dynamic scaling with predictive scaling for better and optimal results.

When capacity planning, identify metrics that impact performance because Auto Scaling creates tracking policies based on the most popular metric for that resource type. And configure CloudWatch to track the identified metrics.

Conclusion

To make the best of AWS Auto Scaling, it is crucial to choose a suitable scaling strategy (or mix them up) according to your application's needs and business goals. In addition, you must know which metrics and constraints are critical for su best to consult an AWS partner who can guide you on how to solve scaling issues for your bu

Si **S Consulting Partner** that can help you customize auto-scaling strategies to suit your bu the right scaling recommendation. So **request a free consultation** with our highly ex expert day

Hiren Dhaduk

Hiren is VP of Technology at Simform with an extensive experience in helping enterprises and startups streamline their business performance through data-driven innovation.

Your comment here*

Sign up for the free Newsletter

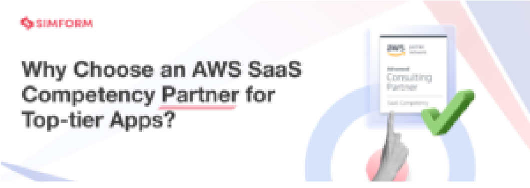
For exclusive strategies not found on the blog

Sign up today!

Related Posts

AWS re:Invent 2023: A Leap in Generative AI and Tech Innovations

Hiren Dhaduk
DECEMBER 15, 2023


SIMFORM
Why Choose an AWS SaaS Competency Partner for Top-tier Apps?
aws
Consulting Partner

Why Choose AWS SaaS Competency Partners for Top-tier Apps?

Hiren Dhaduk
DECEMBER 13, 2023

How does Microservices Architecture Work? A Complete Guide

Hiren Dhaduk
DECEMBER 1, 2023

Areas We Serve

n Francisco
3 Market Street, Suite
00 San Francisco, CA 94111
A

ll us now

Canada
15-115 – Bentall II Vancouver BC,
V7X 1M8

Call us now

India
501 Binori B Sq-2, Nr
DoubleTree By Hilton, Bopal Rd,
Ahmedabad-380054, India

Call us now

Los Angeles
700 S Flower Street, Suite
1000, Los Angeles, CA 90017
USA

Call us now

SIMFORM

★ 5.0 Clutch ★ 4.9 GoodFirms

DMCA PROTECTED

About Simform

- About Us
- Our Partners
- How Simform works
- Areas We Serve
- Careers
- Contact us

Services

- Product Engineering
- Cloud and DevOps Engineering
- Data Engineering
- Quality Engineering
- Talent Solutions

Resources

- Blog
- Ebooks
- Case Studies