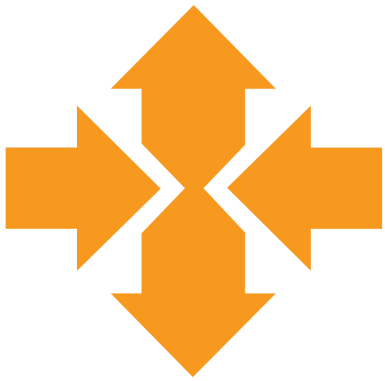
Compute: EC2 Auto Scaling

When to use EC2 Auto Scaling service and it's different features are studied in this lesson. We also present our recommendations on the use of EC2 Auto Scaling service of AWS.

We'll cover the following

Amazon will tell you that Auto Scaling allows you to automatically add or remove EC2 instances based on the fluctuating demands of your application. This sounds great in theory, and while we've certainly seen that work successfully in a few places, it's almost never useful except in very specific situations. You will almost never need to use the auto part of Auto Scaling for the reason it exists.



When to use Auto Scaling?

Let's start by seeing how you should decide how many EC2 instances to run.

Capacity headroom

You obviously need to have enough instances to meet your expected peak demand. But you probably don't want your capacity to exactly match the demand with no leeway. You will want to have some headroom too. This headroom is not waste—it will act as a safety buffer that can absorb many types of unpredictable events. For

example:

- If an availability zone were to go down and you lost half of your instances, the headroom in the remaining instances can compensate for the lost capacity.
- Or if there were to be a sudden increase in demand, the same headroom will be immediately available to take it.
- Or if for some reason the performance of your system were to degrade abruptly (due to a software bug, a bad instance, etc.), that same headroom may help compensate the excess load.

So, a capacity headroom is a wonderful thing. You definitely need some. And if you can afford to, it's probably wise to have a lot of it. It can help you sleep well at night—sometimes in the literal sense.

Cost usage ratio

The main premise of Auto Scaling is that once you decide how much headroom you want, you'll be able to make that headroom a constant size, even as the demand for your instances fluctuates. Therefore, a simpler way to look at Auto Scaling is to see it as just a cost reduction tool. Because what's wrong with having excess headroom during off-peak periods? Absolutely nothing, except cost.

Therefore, the first question you should ask yourself is:

- Are your EC2 costs high enough that any reduction in usage will be materially significant?
- As a thought experiment, consider if your EC2 bill were to go down by 30%—would that be a big deal for your business?

If not, the effort and complexity of getting Auto Scaling working properly is probably not going to be worth it. You might as well just keep the extra headroom during off-peak periods and let it work for you in case of an emergency.

Headroom acts as a safety buffer that can absorb many types of unpredictable events.

Retake Quiz

Demand fluctuations

The other thing to consider is:

- Does your EC2 demand vary enough for Auto Scaling to even matter?

If the fluctuations are not significant, or they are too abrupt, or they are not very smooth, *Auto Scaling will almost certainly not work well for you.*

Nevertheless, in some cases, Auto Scaling can deliver exactly what it says on the tin.

- If you run a business where your EC2 costs are a significant percentage of your expenses and your demand patterns are compatible with Auto Scaling's capabilities, then this can be a handy tool to help you improve your business margins.

OUR RECOMMENDATIONS

↓ OUR RECOMMENDATIONS ↓

You should almost always use **Auto Scaling** if you're using **EC2**! Even if you only have one instance.

Secondary features of Auto Scaling

Auto Scaling has a few secondary features that quite frankly should have been part of EC2 itself.

Replace unhealthy instances

- One of these features is a setting that allows Auto Scaling to automatically replace an instance if it becomes unhealthy.
- If you are already using a load balancer, you can use the same health checks for both the load balancer and Auto Scaling.
- You can also send health check signals using the Auto Scaling API, either directly from your instances (which isn't necessarily a reliable way to send unhealthy signals) or from something that monitors your instances from the outside.

Ease of adding/removing instances

The other nice thing that comes with Auto Scaling is the ability to simply add or remove instances just by updating the desired capacity setting.

- Auto Scaling becomes a launch template for your EC2 instances, and you get a dial that you can turn up or down depending on how many running instances you need.

> There is no faster way to add instances to your fleet than with this method.

One of the limitations of Auto Scaling is that you have to manually remove an instance from the auto scaling group if it becomes unhealthy.

Retake Quiz

In the next lesson, we will take a look at Lambda and it’s usage.