Benefits of Overprovisioning EC2 instances

Consider a scenario where the administrator is designing an ASG for a system which calls for a minimum of 2 instances operating at all times, then common sense would tell us to create an ASG with a minimum of 2 instances. However, in the real world, you will often find that administrators go a step further and have at least 3 instances running at all times.

This might seem odd for the uninitiated, after all, is not the entire point of autoscaling for us to be saving costs by using only the required amount of resources? So then why would we wish to provision more EC2 instances than absolutely necessary?

This is because if an Availability Zone (AZ) outage occurs, your Auto Scaling Group (ASG) will spin up a new instance in an unaffected AZ to compensate for the lost one. However, this new instance doesn't come online immediately, so for a short while, you'll be running with only 1 active instance—leaving your application below the minimum threshold required to operate.

Therefore, over-provisioning by just one instance can be the difference between maintaining your app's availability and leaving it vulnerable during an AZ outage making it a simple, proactive approach that buys you time during unexpected events, and ensures that your application stays resilient and continues to readily meet user demand.