Cloud Instances and Auto-Scaling - Part 3

This lesson discusses the auto-scaling of cloud instances.

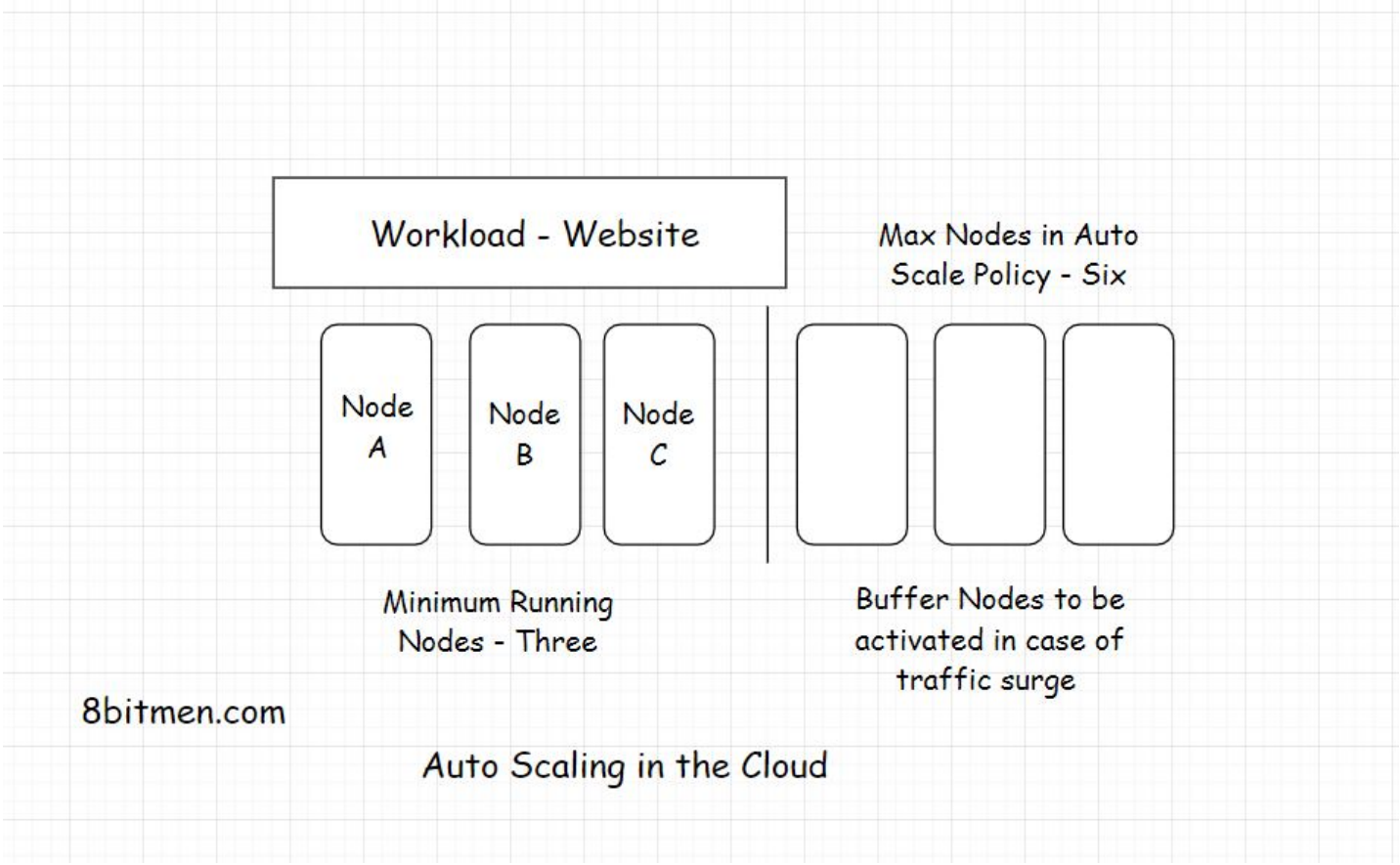
We'll cover the following

What is auto-scaling?

Auto-scaling is the cloud platform's ability to react to the variation in the live traffic load on the workload by spinning instances up and down on the fly.

The ability to auto-scale enables the cloud to augment or reduce the computing power of the instance cluster based on the demand, ensuring smooth handling of the traffic surge and slump. Therefore, when additional instances are added on the fly, they share the load of the already running instances, diminishing the risk of them crumbling under the heavy traffic load.

When the traffic subsides and the workload needs comparatively less computing power, the cloud auto-scaler obliges by freeing instances by shutting them down.



Benefits of auto-scaling

Autoscaling has several benefits. The most important of them is *high availability*.

If a few instances in the cluster go down due to some kind of failure, new instances are automatically spun up to share the load on the cluster. Due to this behavior, the service stays available. This is also known as *fault tolerance*. I'll discuss fault tolerance and high availability in detail later in the course.

Autoscaling provides the ability to the platform to handle node failures smoothly without any human intervention. Another upside of auto-scaling its cost-effectiveness.

When the traffic subsides, additional instances, that were added earlier, are terminated and removed from the fleet. Therefore, the businesses only have to pay for the compute that the service utilizes.

Auto-scaling is also known as *auto-provisioning*. The cloud auto-scaler logic runs on a predefined set of rules and policies. But why do we need to set rules and policies upfront when deploying our workload on the cloud?

Let's find out.

Auto-scaling configurations

Scheduled auto-scaling

The cloud auto-scaler mechanism runs on predefined rules and policies simply due to the reason that businesses, especially startups, have limited resources. They can't just keep adding up server instances on the fly. Computing power costs serious money. We do have to set the rules and configurations as per our budget.

Scheduled auto-scaling is proactive scheduling where we set up all the configuration upfront like the maximum number of instances that can be summoned in the fleet for the support, maximum CPU utilization, and so on.

The scheduled auto-scaling policy holds all the data that commands the auto-scaler on how to react when the workload is hit with a traffic surge. There are various ways to make auto-scaling as effective as possible for services running on the cloud, including *predictive* and *dynamic auto-scaling*.

Predictive auto-scaling

Predictive autoscaling makes use of machine learning to study recent and historical traffic patterns for respective workloads. Based on the study, the right number of instances are provisioned to serve the anticipated traffic.

Dynamic auto-scaling

Dynamic autoscaling is the method where instances are spun up on the fly based on several different metrics such as CPU utilization of the instances, load balancer utilization, and monitoring metrics.

CPU utilization

In the autoscale policy, the threshold for the CPU utilization of the cluster is set, for instance, to 75%, beyond which new instances start spinning up to share the load on the workload.

Load balancer utilization

Another trigger to spin up instances is the requests handled per second by the load balancer. Depending on the value set, an instance can be spun up or terminated.

Monitoring metrics

Besides the above two metrics, monitoring metrics like the container stats, are also

considered when setting up the autoscale policy.

To set an effective auto-scale policy all the above ways are ideally used in conjunction to achieve the best results.

Alright, folks!! With this, we have reached the end of the introductory chapter. Let’s proceed to clusters and how they typically work in the next chapter.

Before we move on, however, there is a brief quiz in the next lesson.