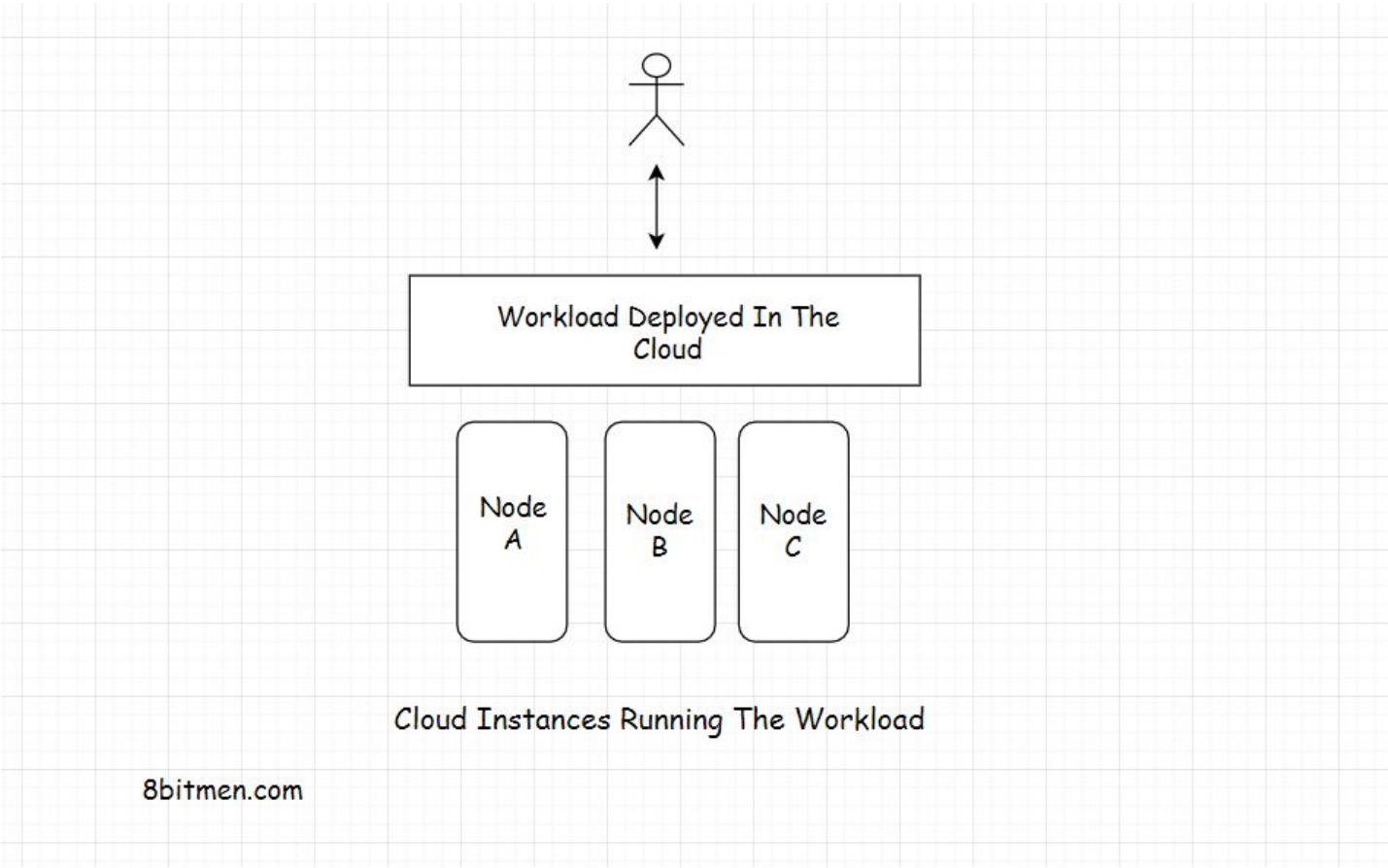
Cloud Instances and Auto-Scaling - Part 1

This lesson provides insight into cloud instances.

We'll cover the following

What is an instance in the cloud?

A server running our application in the cloud is called an *instance*. We can think of one server as one instance. The terms *server* and *instance* can be used interchangeably but in the cloud computing universe, the term instance is preferred over the server. *Instance, server, node, server node*, they all mean the same thing.



Simplifying it even further, imagine a simple _ Create Read Update Delete (CRUD)_ – based app hosted by an *Apache Tomcat* server on our local machine, that is our laptop. Now, when we deploy the app to the cloud, in the cloud computing jargon

laptop. Now, when we deploy the app to the cloud, in the cloud computing jargon that app becomes the *workload* and the *Tomcat* server node becomes an *instance*.

Instances are automatically spun up and down all the time by the cloud platform based on the compute requirements of the workload. This is known as *auto-scaling*.

Instance groups

A workload can be hosted by one single instance or a cluster of instances. They can also be spread out geographically in different availability zones across the globe. Using a cluster of instances enables us to make our system *fault-tolerant* and achieve *high availability*. More on that later in this course. Instances are often grouped in a cluster and referred to as an *instance group*.

Instance groups enable us to enforce common policies, rules, and configuration across multiple instances. These policies can be configured on high availability, auto-healing, load balancing, applying updates, and so on. Overall these instance groups make the instance management easier.

Instances have no dependency on the workload and vice versa. Both are loosely coupled. This allows them to spin up and down according to demand. Also, having no dependency makes the workload portable. It can be moved around from on-prem to cloud and across different cloud platforms without much hassle.

Generally, instances in the cloud are *virtual machines* that run the images of operating systems like *Linux* and *Windows*. We can also create custom images based on our requirements or import existing images and run them on our cloud instances.

Instances can also run containers with container optimized operating systems. Using instance templates, we can clone new instances with the existing configuration eventually saving a lot of time.

If you are unaware of terms like *containers or virtual machines*, no worries. I'll discuss them in detail in future lessons. For now, let's focus on instances.

When deploying our workload on the cloud, we can choose the hardware properties of our instances, such as the number of virtual CPUs, memory, storage capacity, and so on. Instances are categorized into several different types by the cloud providers where each category of instances serves certain specific use cases.

In the next lesson, let’s find out what these types of instances are.