Cloud Workload

This lesson introduces workload and its types in the cloud.

We'll cover the following

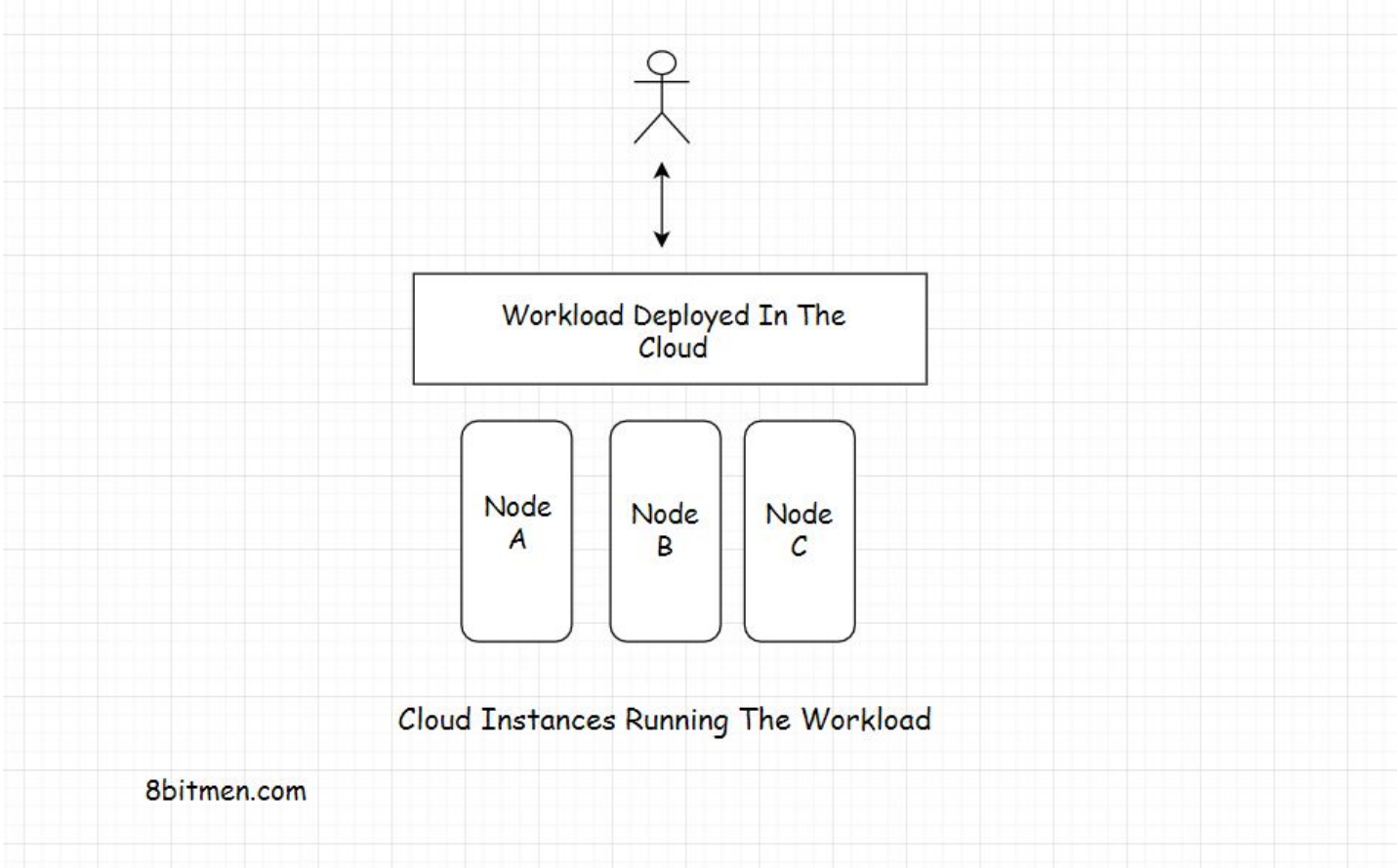
- What is a workload in the cloud?
- Types of workloads in the cloud
- Workloads classified by resource requirements
  - General compute
  - CPU intensive
  - Memory intensive
  - GPU accelerated computation
  - Storage optimized database workloads
- Workloads classified by user traffic patterns
  - Static workloads
  - Periodic workloads
  - Unpredictable workloads
  - Hybrid workloads

What is a workload in the cloud?

A service deployed on the cloud is called a *workload*. That service could be one modest monolith or a massive one composed of hundreds of microservices working in conjunction with each other. The term workload signifies *abstraction* and *portability*.

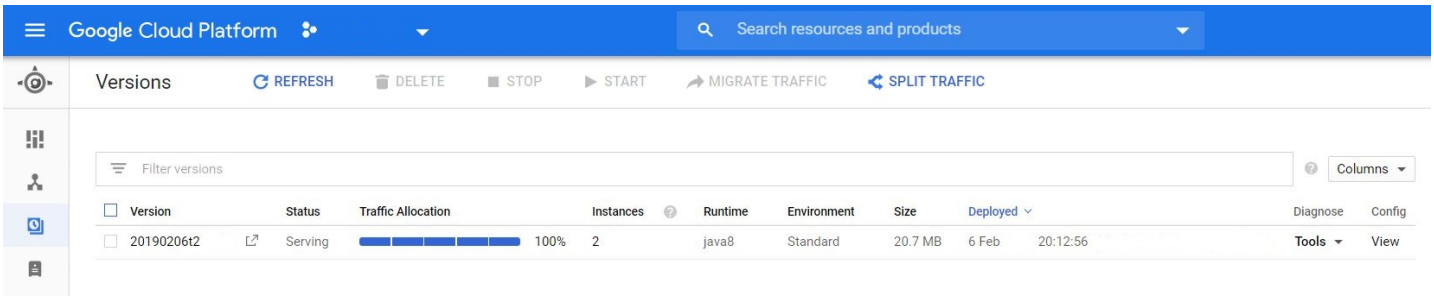
It can be moved around, across different cloud platforms or from on-prem to the cloud and vice-versa, without any dependencies or much hassle. This portability of workloads is primarily facilitated by the *container technology*. We'll talk about containers in more detail later in this course.

The diagram below shows a workload deployed on the cloud run by multiple machines.



Therefore every time we deploy our workload on the cloud, the platform creates a new version of it with a *version id*. Having different versions of the same workload helps with the A/B testing. We can switch between different versions, split traffic between the multiple versions based on our requirements, and more.

Below is a snapshot of another service that I run on Google Cloud. You can see the workload versions, status, traffic allocation, instances assigned, runtime, workload size, etc.



Types of workloads in the cloud

Workloads can be classified into different categories primarily based on their resource requirements and traffic patterns.

Let's take a look at the workloads classified by the resource requirements.

Workloads classified by resource requirements

General compute

Workloads that require general computing power are general compute workloads. They are typically web applications, containerized microservices, and so on. They do not have any specific computational requirements and are easily run using the resources that are provisioned by default by the cloud provider.

CPU intensive

CPU intensive workloads have high computational requirements. They need powerful servers to run. These kinds of workloads are deep learning applications, highly scalable multiplayer gaming apps that are expected to handle a large number of concurrent users, Big Data analytics services, 3D modeling, video encoding, etc.

Memory intensive

Memory intensive workloads need large CPU memory to execute tasks. These are typically distributed databases, caches, real-time streaming, data analytics, and so on.

GPU accelerated computation

Workload running processes, such as seismic analysis, computational fluid dynamics, autonomous vehicle algorithms, speech recognition, etc., require the power of GPUs along with the CPUs to run accelerated tasks. These are known as the GPU accelerated computation workloads.

Storage optimized database workloads

These workloads are primarily highly scalable, e.g., *NoSQL* databases, in-memory databases, data warehouses, etc. These workloads are designed to store large amounts of data in an optimized fashion.

Now, let's have a look at the workloads classified by the traffic patterns.

Workloads classified by user traffic patterns

Static workloads

Static workloads always utilize resources within a certain range. There are no

surprises, they receive no traffic bursts. These workloads can be a utility deployed on the cloud and accessed by a limited number of users in a private network. For instance, it can be an organization-wide tax-calculation utility or a knowledge base on something.

Periodic workloads

These workloads are triggered only at specific times and for a specific duration, like for a few days in a month or a week. The electricity bill payment app, when accessed by users only during the last few days of the month, is a good example of this.

Serverless compute works best for these kinds of applications, where there is no need to pay for idle instances, only the compute utilized.

Unpredictable workloads

These workloads include popular apps like social networks, online multiplayer games, game streaming apps, and so on. Traffic on these services can spike exponentially at any time. Social networks experience traffic surges all the time during big global or countrywide events.

Hybrid workloads

Hybrid workloads as the name implies is the hybrid or the mix of all the above types of workloads.

In the next lesson, let’s discuss the instances in the cloud.