Health Checks

We mentioned Health Checks in the last section, but what are they exactly and what purpose do they serve? Well that is what we will discuss in this section.

Health Checks as the name suggests, are processes that monitor the status and performance of our AWS resources and components to determine whether or not they are functioning correctly, i.e. whether or not they are “Healthy”. As mentioned in the last section, they are an integral part of load balancers and are often used to assess whether backend assets (such as EC2 instances) are capable of handling traffic. They are quite widely used and help us better our systems by providing:

- **High Availability**: Health checks help maintain high availability by ensuring that traffic is only routed to instances that are operational. If an instance fails a health check, it is marked as unhealthy, and the load balancer stops sending traffic to it.
- **Fault Tolerance**: By identifying unhealthy instances quickly, health checks facilitate swift recovery and failover, allowing the system to redirect traffic to healthy instances and minimize downtime.
- **Performance Monitoring**: Health checks can provide insights into the performance of application components, allowing teams to detect and address issues proactively.

Of the three types of load balancers, Gateway Load Balancers do not support health checks leaving us with the other two, namely: Application Load Balancers and Network Load Balancers.

Application Load Balancers support HTTP health checks and are the best at it, allowing them to check specific API URLs and endpoints, directing the requests only to the paths that it has determined to be capable of performing work successfully.

Network Load Balancers on the other hand, primarily handle health checks related to TCP/UDP traffic, and are especially useful for operations requiring low latency and high throughput, such as real-time gaming or IoT applications.

Note however that Network Load Balancers are also capable of performing HTTP health checks, though NLB health checks are simpler in nature, typically only checking if a specific port is open for connection, and not performing anything beyond that making them extremely suitable for simple services that do not require complex health verification.

Finally, an often overlooked aspect of health checks in AWS are their integrations with other features such as Auto Scaling. Not many know this but we can actually use AWS Auto Scaling to automatically add or remove instances based on health check status.

For example, if an instance is marked as unhealthy, Auto Scaling can launch a new instance to replace it, automating maintenance and ensuring continuous availability. This is usually done by utilizing a sub-feature of Auto Scaling in AWS called Auto Scaling actions. I shall spare the reader with a more in-depth explanation of the sub-feature as I judge it to be beyond the scope of both this section and this book in general.