https://www.youtube.com/c/TariqKhan

Core Cloud Services - Azure Architecture and Service guarantees

In this module, you will:

- Explore the physical structure of Cloud infrastructure
- Understand the service level agreements provided by Azure
- Learn how to provide your own service level agreements for your apps

Understand Datacenters and Regions in Cloud

What is a region?

A **region** is a geographical area on the planet containing at least one, but potentially multiple datacenters that are nearby and networked together with a low-latency network. Azure intelligently assigns and controls the resources within each region to ensure workloads are appropriately balanced.

When you deploy a resource in Azure, you will often need to choose the region where you want your resource deployed.

Understand Geographies in Azure Cloud

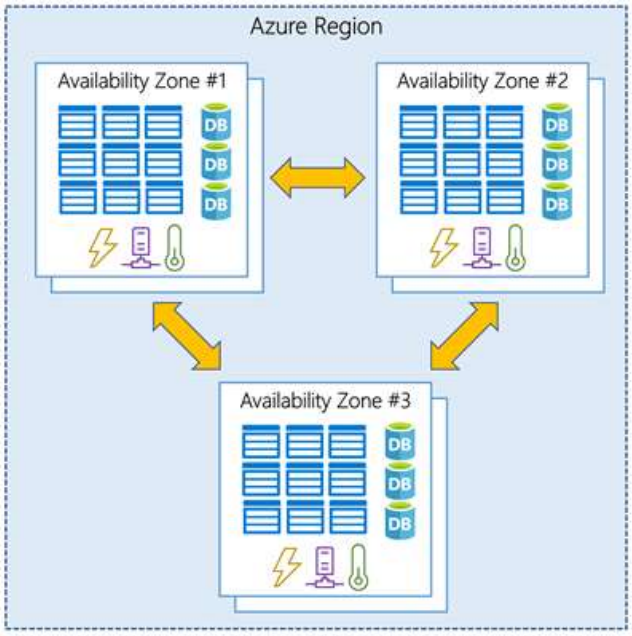
Azure divides the world into *geographies* that are defined by geopolitical boundaries or country borders. An Azure geography is a discrete market typically containing two or more regions that preserve data residency and compliance boundaries. This division has several benefits.

- Geographies allow customers with specific data residency and compliance needs to keep their data and applications close.
- Geographies ensure that data residency, sovereignty, compliance, and resiliency requirements are honored within geographical boundaries.
- Geographies are fault-tolerant to withstand complete region failure through their connection to dedicated high-capacity networking infrastructure.

What is an Availability Zone?

Availability Zones are physically separate datacenters within an Azure region.

Each Availability Zone is made up of one or more datacenters equipped with independent power, cooling, and networking. It is set up to be an *isolation boundary*. If one zone goes down, the other continues working. Availability Zones are connected through high-speed, private fiber-optic networks.



What is a region pair?

Each Azure region is always paired with another region within the same geography (such as US, Europe, or Asia) at least **300 miles away**. This approach allows for the replication of resources (such as virtual machine storage) across a geography that helps reduce the likelihood of interruptions due to events such as natural disasters, civil unrest, power outages, or physical network outages affecting both regions at once. If a region in a pair was affected by a natural disaster, for instance, services would automatically fail over to the other region in its region pair.

Examples of region pairs in Azure are West US paired with East US, and SouthEast Asia paired with East Asia.

SLAs for Azure products and services

There are three key characteristics of SLAs for Azure products and services:

1. Performance Targets
2. Uptime and Connectivity Guarantees
3. Service credits

Performance Targets

An SLA defines performance targets for an Azure product or service. The performance targets that an SLA defines are specific to each Azure product and service. For example, performance targets for some Azure services are expressed as uptime guarantees or connectivity rates.

Uptime and Connectivity Guarantees

A typical SLA specifies performance-target commitments that range from 99.9 percent ("three nines") to 99.999 percent ("five nines"), for each corresponding Azure product or service. These targets can apply to such performance criteria as uptime or response times for services.

The following table lists the potential cumulative downtime for various SLA levels over different durations:

UPTIME AND CONNECTIVITY GUARANTEES			
SLA %	**Downtime per week**	**Downtime per month**	**Downtime per year**
99	1.68 hours	7.2 hours	3.65 days
99.9	10.1 minutes	43.2 minutes	8.76 hours
99.95	5 minutes	21.6 minutes	4.38 hours
99.99	1.01 minutes	4.32 minutes	52.56 minutes
99.999	6 seconds	25.9 seconds	5.26 minutes

For example, the SLA for the Azure Cosmos DB (Database) service SLA offers 99.999 percent uptime, which includes low-latency commitments of less than 10 milliseconds on DB read operations as well as on DB write operations.

Service Credits

SLAs also describe how Microsoft will respond if an Azure product or service fails to perform to its governing SLA's specification.

For example, customers may have a discount applied to their Azure bill, as compensation for an under-performing Azure product or service. The table below explains this example in more detail.

The first column in the table below shows monthly uptime percentage SLA targets for a single instance Azure Virtual Machine. The second column shows the corresponding service credit amount you receive if the *actual* uptime is less than the specified SLA target for that month.

SERVICE CREDITS	
MONTHLY UPTIME PERCENTAGE	**SERVICE CREDIT PERCENTAGE**
< 99.9	10
< 99	25
< 95	100

https://www.youtube.com/c/TariqKhan