

Azure Architecture and Service Guarantees

1 Datacenters and Regions

Definition: Region

A geographical area on the planet containing at least one, but potentially multiple datacenters networked together

Note that some services or features are only available in certain regions

1.1 Special Regions

- US DoD and Gov are separate datacenters specifically for government applications
- China datacenters are managed by 21Vianet

2 Geographies

Geographies are defined by geopolitical boundaries or county borders. This has the benefits:

- Allow customers with specific compliance to keep their data close
- Ensure data residency, sovereignty, compliance and resiliency requirements are honoured within geographical boundaries.
- Ensure fault tolerance.

Definition: Data residency

The physical or geographic location of an organization's data or information.

Azure has the following geographies:

- Americas
- Europe
- Asia Pacific
- Middle East and Africa

3 Availability Zones

Definition: Availability Zone

Physically separate datacenters within an Azure region

This means that if one zone goes down, the other continues working

3.1 Using Availability Zones

Azure services that support availability zones fall in two categories:

- Zonal services - Pin the resource to a specific zone
- Zone-redundant services - Platform replicates automatically across zones

4 Region Pairs

Definition: Region Pair

Another region within the same geography at least 300 miles away

This reduces the likelihood of interruptions such as natural disasters.

Important

An availability zone sits within a region pair. So an availability zone may be two datacenters in the UK, a region pair may be one in the UK and one in France

Advantages of region pairs:

- If there's a large azure outage, one region out of every pair is prioritized to be restored as quickly as possible
- Planned Azure updates are rolled out to paired regions one region at a time to minimise downtime and risk of application outage
- Data continues to reside in the same geography as its pair for compliance purposes

5 Service Level Agreements

Important

Azure does not provide SLAs for most services under the Free or Shared tiers

5.1 Performance targets

An SLA provides performance targets for a product or service

5.2 Uptime and connectivity guarantees

Typical SLA ranges from 99.9% to 99.999%

5.3 Service credits

Define how Microsoft will respond if a product fails to perform to its SLA spec.

The lower the uptime percentage, the greater the service credit percentage given

5.4 Composing SLAs across services

Calculating the SLA of two services together is just multiplying them

However if you install redundancy such as two databases, the SLA is calculated with

$$1 - (\text{Failure chance of service 1} \times \text{Failure chance of service 2})$$

And this result is then multiplied by the SLA of the other component to get the SLA

5.5 Improving App Reliability

Definition: Resiliency

The ability of a system to recover from failures and continue to function

Definition: Failure Mode Analysis

Identifying points of failure and defining how the application will respond to those failures

Definition: Availability

The time that the system is functional and working

Maximizing availability requires implementing measures to prevent service failures. However this can be difficult and expensive.

You need to ensure that the minimum SLA in your system is one that you are happy with your service performing at.