AZ-900 Revision Chapter – 4: Understanding Azure pricing and Support

Skill 4.4: Describe Azure service level agreements

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

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

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.

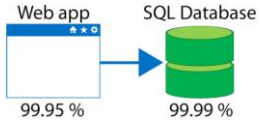
Service Credits

SERVICE CREDITS

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

Calculating downtime

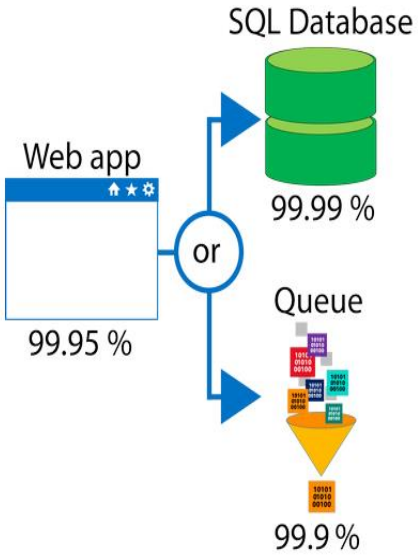
Consider an App Service web app that writes to Azure SQL Database. These Azure services currently have the following SLAs:



In this example, if either service fails the whole application will fail. In general, the individual probability values for each service are independent. However, the composite SLA value for this application is:
99.95 percent × 99.99 percent = 99.94 percent

This means the **combined probability of failure** is higher than the individual SLA values. This isn't surprising, because an application that relies on multiple services has more potential failure points.

Conversely, you can improve the composite SLA by creating independent fallback paths. For example, if the SQL Database is unavailable, you can put transactions into a queue for processing at a later time.



With this design, the application is still available even if it can't connect to the database. However, it fails if both the database and the queue fail simultaneously.
If the expected percentage of time for a simultaneous failure is **0.0001 × 0.001**, the composite SLA for this combined path of a database or queue would be:
1.0 − (0.0001 × 0.001) = 99.99999 percent
Therefore, if we add the queue to our web app, the total composite SLA is:
99.95 percent × 99.99999 percent = ~99.95 percent
99.96

Resiliency

Resiliency is the ability of a system to recover from failures and continue to function. It's not about avoiding failures, but responding to failures in a way that avoids downtime or data loss. The goal of resiliency is to return the application to a fully functioning state following a failure. High availability and disaster recovery are two crucial components of resiliency.

When designing your architecture you need to design for resiliency, and you should perform a Failure Mode Analysis (FMA). The goal of an FMA is to identify possible points of failure and to define how the application will respond to those failures.