**Azure Traffic Manager**

The Azure Traffic Manager service is a DNS-based traffic load balancer that distributes traffic across services that are distributed across different Azure regions.

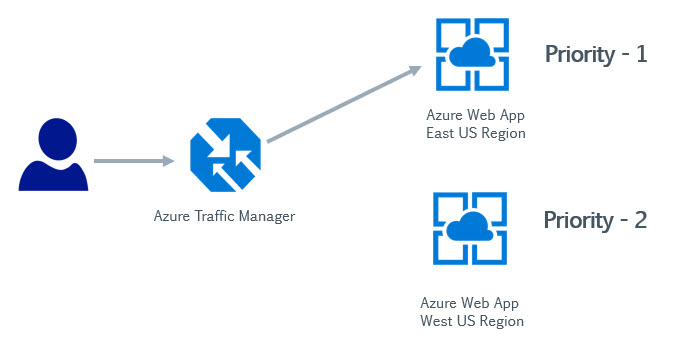
The Traffic Manager service is used to direct client requests to the most appropriate service endpoint that is based on a traffic-routing method and the health of the endpoints.

The different traffic routing methods available for the Azure Traffic Manager are

* Priority – Route traffic to another endpoint in case the primary fails.
* Weighted – Route traffic to different endpoints based on weight.
* Performance - you want end users to use the "closest" endpoint in terms of the lowest network latency.
* Geographic - geographic location their DNS query originates from.
* Multivalue – Here different endpoints are sent to the client. The client then selects the endpoint to send the request to.
* Subnet – This maps a set of end-user IP address ranges to a specific endpoint within a Traffic Manager profile.

Below is an example of the Priority routing method that can be used with the Azure Traffic Manager service

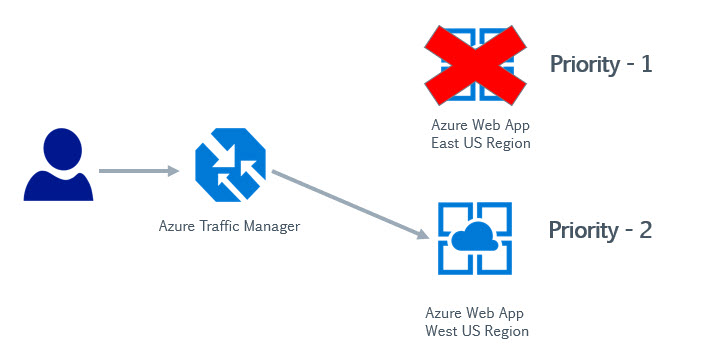
Here we are assuming that a company has similar web applications , both are running using the Azure Web App service. One web application is running in the East US Region and the other is running in the West US Region.



1. Here we create a Traffic Manager profile and create two endpoints. Each endpoint points to each Azure Web app respectively. We assign a priority of 1 to the service endpoint attached to the Azure Web App running in the East US region and  a priority of 2 to the other service endpoint.

1. Here users would make requests to the Traffic Manager service.

2. The requests could be initially be directed to an Azure Web App located in the East US region , since there is a priority of 1 to the service endpoint attached to this endpoint.

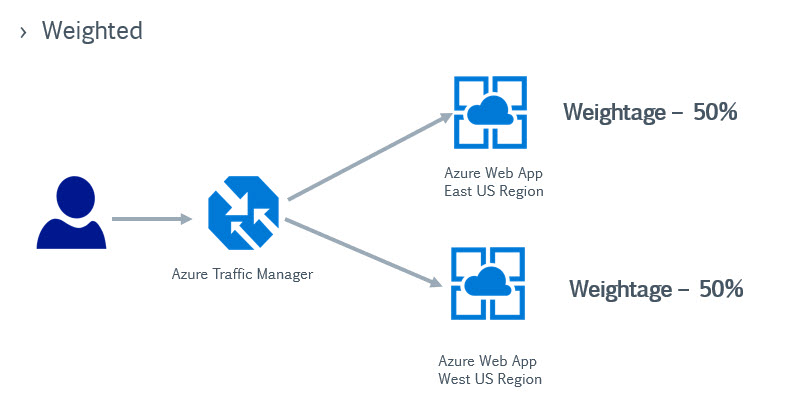


3. Now let's say there is an issue with the web application running in the East US region, Azure Traffic Manager would then understand that there is an issue with the web application running in this region.

It would then start redirecting user requests to the second endpoint which has the Priority of 2.

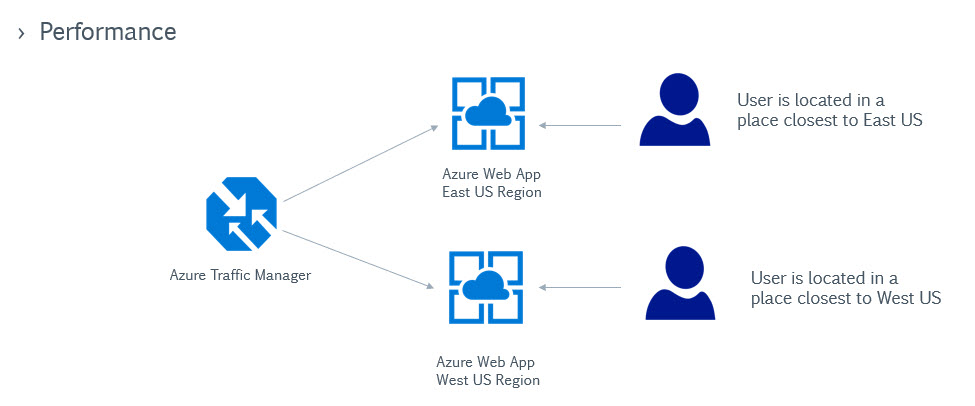
Hence over here you are adding a higher availability to your architecture by ensuring that user requests are always adhered to by redirecting requests if the primary service fails for any reason.

If you use the Weighted Routing method , you can actually load balance requests across multiple service endpoints



Over here , users requests would be directed or load balanced across both web applications running in different regions.

In the Performance routing method as shown below, users will be directed based on the least latency of an endpoint.



And then we have the Geographic routing method wherein users would be directed to an endpoint based on their geographic location

