Mustafa Saifee(Microsoft
Student Partner 2015-19)

AZURE FUNDAMENTALS

Core Cloud Services - Azure networking options

-Microsoft Learning

Contents

Introduction

You just started working at a startup that's fundamentally disrupting the vitamin industry with simple customization and affordable monthly subscriptions. While business is booming on the e-commerce site, your data center is starting to struggle to keep up with user demand. Your service fails when too many users sign in at the same time, and you're facing more scheduled and unscheduled maintenance windows than you'd like.

Your site is based in Silicon Valley, so you also find that a network delay is especially bad for users who are located in other regions, such as Europe and Asia.

Therefore, you convince your team to move the site to the cloud to help save costs. But how can Azure, specifically, help your site run better?

As it turns out, managing networks on Azure isn't entirely different from managing on-premises networks. Let's discover why.

Learning objectives

In this module, you will learn:

- How an Azure virtual network provides secure network communication among resources such as virtual machines and other networks
- What high availability and resiliency mean and how Azure Load Balancer can increase resiliency within a single geographic region
- What latency is and how Traffic Manager helps reduce network latency and provides resiliency across geographic locations

Deploy your site to Azure

Your first step will likely be to re-create your on-premises configuration in the cloud.

This basic configuration will give you a sense of how networks are configured, and how network traffic moves in and out of Azure.

Your e-commerce site at a glance

Larger enterprise systems are often composed of multiple inter-connected applications and services that work together. You might have a front-end web system that displays inventory and allows customers to create an order. That might talk to a variety of web services to provide the inventory data, manage user profiles, process credit cards, and request fulfillment of processed orders.

There are several strategies and patterns employed by software architects and designers to make these complex systems easier to design, build, manage, and maintain. Let's look at a few of them, starting with *loosely coupled architectures*.

Benefits of Loosely Coupled Architectures

Using an N-tier architecture

An architectural pattern that can be used to build loosely coupled systems is *N-tier*.

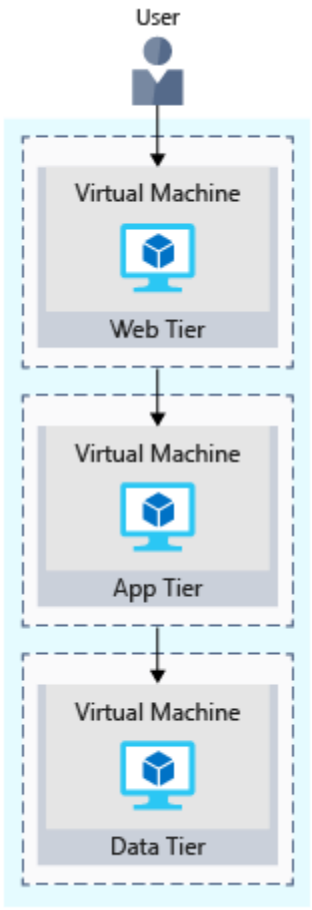
An N-tier architecture divides an application into two or more logical tiers. Architecturally, a higher tier can access services from a lower tier, but a lower tier should never access a higher tier.

Tiers help separate concerns and are ideally designed to be reusable. Using a tiered architecture also simplifies maintenance. Tiers can be updated or replaced independently, and new tiers can be inserted if needed.

Three-tier refers to an n-tier application that has three tiers. Your e-commerce web application follows this three-tier architecture:

- The **web tier** provides the web interface to your users through a browser.
- The **application tier** runs business logic.
- The **data tier** includes databases and other storage that hold product information and customer orders.

The following illustration shows the flow of a request from the user to the data tier.

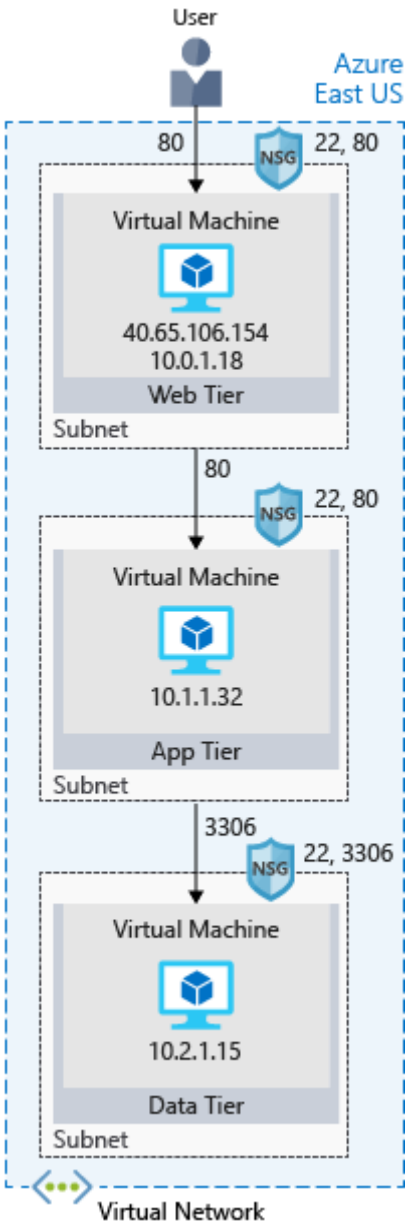


When the user clicks the button to place the order, the request is sent to the web tier, along with the user's address and payment information. The web tier passes this information to the application tier, which would validate payment information and check inventory. The application tier might then store the order in the data tier, to be picked up later for fulfillment.

Your e-commerce site running on Azure

Azure provides many different ways to host your web applications, from fully pre-configured environments that host your code, to virtual machines that you configure, customize, and manage.

Let's say you choose to run your e-commerce site on virtual machines. Here's what that might look like in your test environment running on Azure. The following illustration shows a three-tier architecture running on virtual machines with security features enabled to restrict inbound requests.

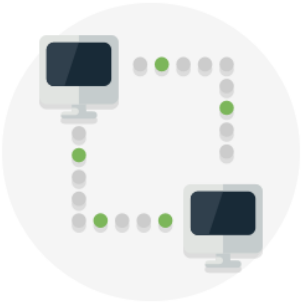


Let's break this down.

What's an Azure region?

A *region* is one or more Azure data centers within a specific geographic location. East US, West US, and North Europe are examples of regions. In this instance, you see that the application is running in the East US region.

What's a virtual network?



A *virtual network* is a logically isolated network on Azure. Azure virtual networks will be familiar to you if you've set up networks on Hyper-V, VMware, or even on other public clouds. A virtual network allows Azure resources to securely communicate with each other, the internet, and on-premises networks. A virtual network is scoped to a single region; however, multiple virtual networks from different regions can be connected together using virtual network peering.

Virtual networks can be segmented into one or more *subnets*. Subnets help you organize and secure your resources in discrete sections. The web, application, and data tiers each have a single VM. All three VMs are in the same virtual network but are in separate subnets.

Users interact with the web tier directly, so that VM has a public IP address along with a private IP address. Users don't interact with the application or data tiers, so these VMs each have a private IP address only.

You can also keep your service or data tiers in your on-premises network, placing your web tier into the cloud, but keeping tight control over other aspects of your application. A *VPN gateway* (or virtual network gateway), enables this scenario. It can provide a secure connection between an Azure Virtual Network and an on-premises location over the internet.

Azure manages the physical hardware for you. You configure virtual networks and gateways through software, which enables you to treat a virtual network just like your own network. You choose which networks your virtual network can reach, whether that's the public internet or other networks in the private IP address space.

What's a network security group?

A *network security group*, or NSG, allows or denies inbound network traffic to your Azure resources. Think of a network security group as a cloud-level firewall for your network.

For example, notice that the VM in the web tier allows inbound traffic on ports 22 (SSH) and 80 (HTTP). This VM's network security group allows inbound traffic over these ports from all sources. You can configure a network security group to accept traffic only from known sources, such as IP addresses that you trust.

Note

Port 22 enables you to connect directly to Linux systems over SSH. Here we show port 22 open for learning purposes. In practice, you might configure VPN access to your virtual network to increase security.

Summary

Your three-tier application is now running on Azure in the East US region. A *region* is an Azure data center within a specific geographic location.

Each tier can access services only from a lower tier. The VM running in the web tier has a public IP address because it receives traffic from the internet. The VMs in the lower tiers, the application and data tiers, each have private IP addresses because they don't communicate directly over the internet.

Virtual networks enable you to group and isolate related systems. You define *network security groups* to control what traffic can flow through a virtual network.

The configuration you saw here is a good start. But when you deploy your e-commerce site to production in the cloud, you'll likely run into the same problems as you did in your on-premises deployment.

Scale with Azure Load Balancer

You now have your site up and running on Azure. But how can you help ensure your site is running 24/7?

For instance, what happens when you need to do weekly maintenance? Your service will still be unavailable during your maintenance window. And because your site reaches users all over the world, there's no good time to take down your systems for maintenance. You may also run into performance issues if too many users connect at the same time.

What are availability and high availability?

Availability refers to how long your service is up and running without interruption. *High availability*, or *highly available*, refers to a service that's up and running for a long period of time.

You know how frustrating it is when you can't access the information you need. Think of a social media or news site that you visit daily. Can you always access the site, or do you often see error messages like "503 Service Unavailable"?

You may have heard terms like "five nines availability." Five nines availability means that the service is guaranteed to be running 99.999 percent of the time. Although it's difficult to achieve 100 percent availability, many teams strive for at least five nines.

What is resiliency?

Resiliency refers to a system's ability to stay operational during abnormal conditions.

These conditions include:

- Natural disasters
- System maintenance, both planned and unplanned, including software updates and security patches.
- Spikes in traffic to your site
- Threats made by malicious parties, such as distributed denial of service, or DDoS, attacks

Imagine your marketing team wants to have a flash sale to promote a new line of vitamin supplements. You might expect a huge spike in traffic during this time. This spike could overwhelm your processing system, causing it to slow down or halt, disappointing your users. You may have experienced this disappointment for yourself. Have you ever tried to access an online sale only to find the website wasn't responding?

What is a load balancer?

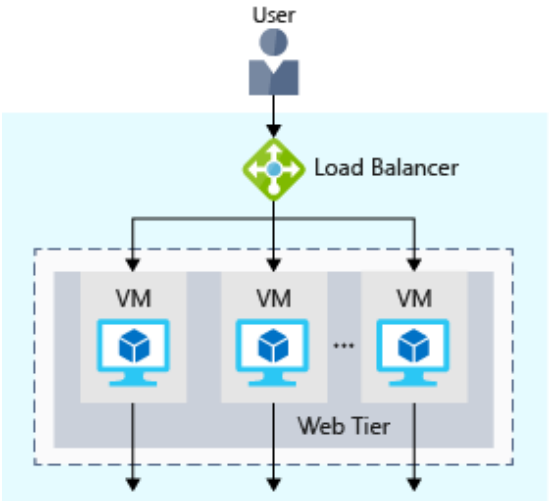
A *load balancer* distributes traffic evenly among each system in a pool. A load balancer can help you achieve both high availability and resiliency.

Say you start by adding additional VMs, each configured identically, to each tier. The idea is to have additional systems ready, in case one goes down, or is serving too many users at the same time.

The problem here is that each VM would have its own IP address. Plus, you don't have a way to distribute traffic in case one system goes down or is busy. How do you connect your VMs so that they appear to the user as one system?

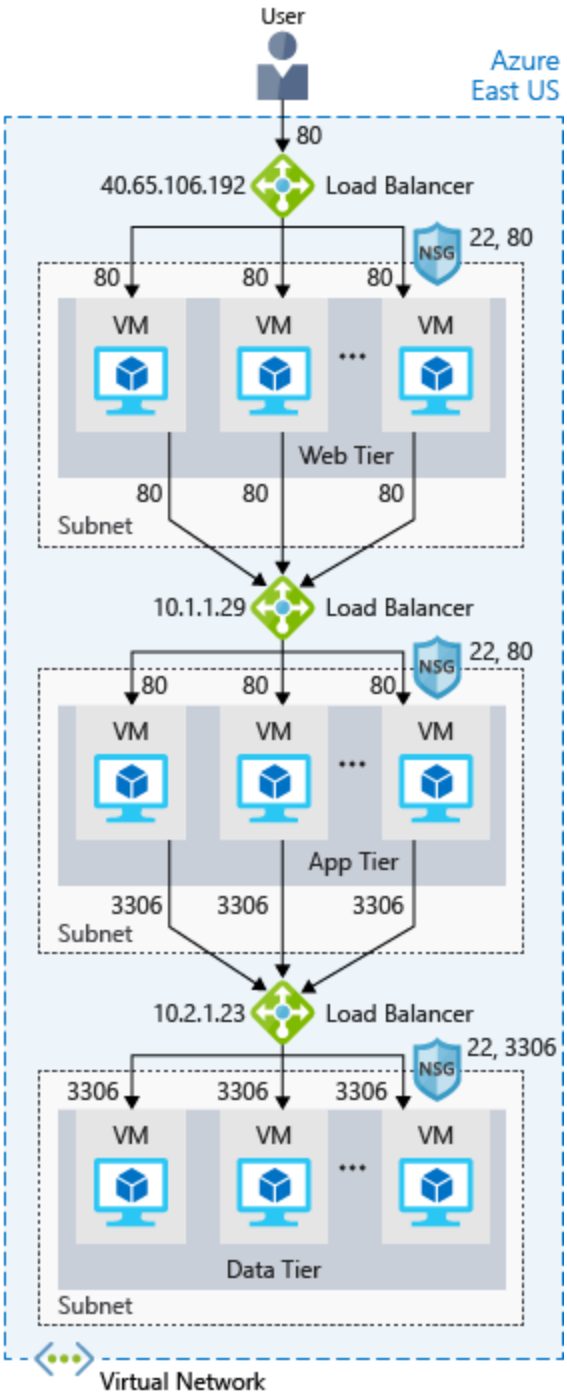
The answer is to use a *load balancer* to distribute traffic. The load balancer becomes the entry point to the user. The user doesn't know (or need to know) which system the load balancer chooses to receive the request.

The following illustration shows the role of a load balancer.



The load balancer receives the user's request and directs the request to one of the VMs in the web tier. If a VM is unavailable or stops responding, the load balancer stops sending traffic to it. The load balancer then directs traffic to one of the responsive servers.

Load balancing enables you to run maintenance tasks without interrupting service. For example, you can stagger the maintenance window for each VM. During the maintenance window, the load balancer detects that the VM is unresponsive, and directs traffic to other VMs in the pool.



For your e-commerce site, the app and data tiers can also have a load balancer. It all depends on what your service requires.

What is Azure Load Balancer?

Azure Load Balancer is a load balancer service that Microsoft provides that helps take care of the maintenance for you. Load Balancer supports inbound and outbound scenarios, provides low latency and high throughput, and scales up to millions of flows for all Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) applications. You can use Load Balancer with incoming internet traffic, internal traffic across Azure services, port forwarding for specific traffic, or outbound connectivity for VMs in your virtual network.

When you manually configure typical load balancer software on a virtual machine, there's a downside: you now have an additional system that you need to maintain. If your load balancer goes down or needs routine maintenance, you're back to your original problem.

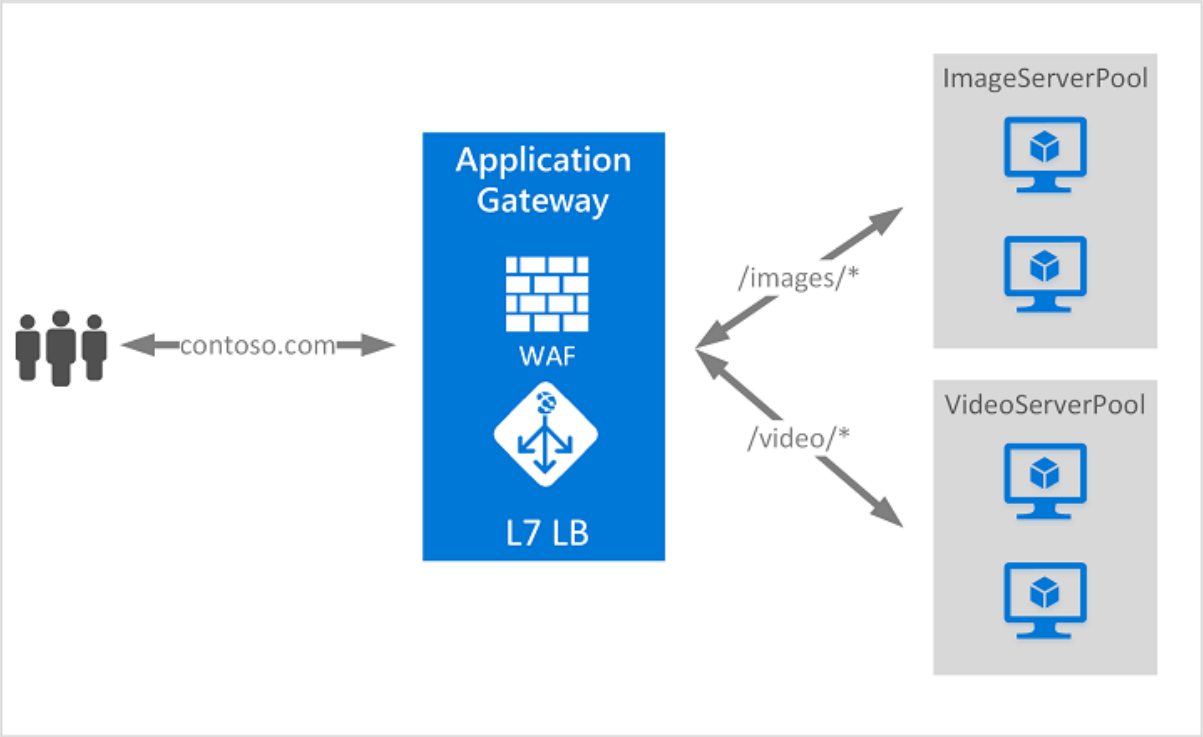
If instead, however, you use Azure Load Balancer, there's no infrastructure or software for you to maintain. You define the forwarding rules based on the source IP and port to a set of destination IP/ports.

The following illustration shows the role of Azure load balancers in a multi-tier architecture.

Azure Application Gateway

If all your traffic is HTTP, a potentially better option is to use Azure Application Gateway. Application Gateway is a load balancer designed for web applications. It

uses Azure Load Balancer at the transport level (TCP) and applies sophisticated URL-based routing rules to support several advanced scenarios.



This type of routing is known as application layer (OSI layer 7) load balancing since it understands the structure of the HTTP message.

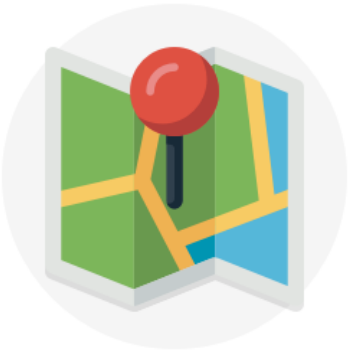
Here are some of the benefits of using Azure Application Gateway over a simple load balancer:

- **Cookie affinity**. Useful when you want to keep a user session on the same backend server.
- **SSL termination**. Application Gateway can manage your SSL certificates and pass unencrypted traffic to the backend servers to avoid encryption/decryption overhead. It also supports full end-to-end encryption for applications that require that.
- **Web application firewall**. Application gateway supports a sophisticated firewall (WAF) with detailed monitoring and logging to detect malicious attacks against your network infrastructure.
- **URL rule-based routes**. Application Gateway allows you to route traffic based on URL patterns, source IP address and port to destination IP address and port. This is helpful when setting up a *content delivery network*.
- **Rewrite HTTP headers**. You can add or remove information from the inbound and outbound HTTP headers of each request to enable important security scenarios, or scrub sensitive information such as server names.

What is a Content Delivery Network?

A content delivery network (CDN) is a distributed network of servers that can efficiently deliver web content to users. It is a way to get content to users in their local region to minimize latency. CDN can be hosted in Azure or any other location. You can cache content at strategically placed physical nodes across the world and provide better performance to end users. Typical usage scenarios include web applications containing multimedia content, a product launch event in a particular region, or any event where you expect a high-bandwidth requirement in a region.

What about DNS?

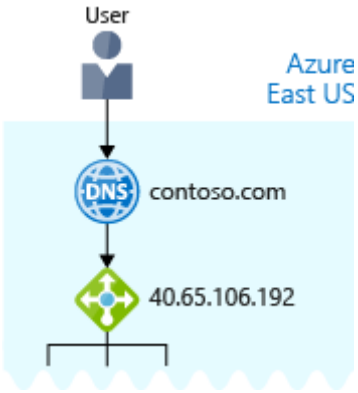


DNS, or Domain Name System, is a way to map user-friendly names to their IP addresses. You can think of DNS as the phonebook of the internet.

For example, your domain name, contoso.com, might map to the IP address of the load balancer at the web tier, 40.65.106.192.

You can bring your own DNS server or use Azure DNS, a hosting service for DNS domains that runs on Azure infrastructure.

The following illustration shows Azure DNS. When the user navigates to **contoso.com**, Azure DNS routes traffic to the load balancer.



Summary

With load balancing in place, your e-commerce site is now more highly available and resilient. When you perform maintenance or receive an uptick in traffic, your load balancer can distribute traffic to another available system.

Although you can configure your own load balancer on a VM, Azure Load Balancer reduces upkeep because there's no infrastructure or software to maintain.

DNS maps user-friendly names to their IP addresses, much like how a phonebook maps names of people or businesses to phone numbers. You can bring your own DNS server, or use Azure DNS.

Reduce latency with Azure Traffic Manager

Previously, you saw how **Azure Load Balancer** helps you achieve high availability and minimize downtime.

Although your e-commerce site is more highly available, it doesn't solve the issue of latency or create resiliency across geographic regions.

How can you make your site, which is located in the United States, load faster for users located in Europe or Asia?

What is network latency?

Latency refers to the time it takes for data to travel over the network. Latency is typically measured in milliseconds.

Compare latency to bandwidth. Bandwidth refers to the amount of data that can fit on the connection. Latency refers to the time it takes for that data to reach its destination.

Factors such as the type of connection you use and how your application is designed can affect latency. But perhaps the biggest factor is distance.

Think about your e-commerce site on Azure, which is in the East US region. It would typically take less time to transfer data to Atlanta (a distance of around 400 miles) than to transfer data to London (a distance of around 4,000 miles).

Your e-commerce site delivers standard HTML, CSS, JavaScript, and images. The network latency for many files can add up. How can you reduce latency for users located far away geographically?

Scale out to different regions

Recall that Azure provides data centers in regions across the globe.

Think about the cost of building a data center. Equipment costs aren't the only factor. You need to provide the power, cooling, and personnel to keep your systems running at each location. It might be prohibitively expensive to replicate your entire data center. But doing so with Azure can cost much less, because Azure already has the equipment and personnel in place.

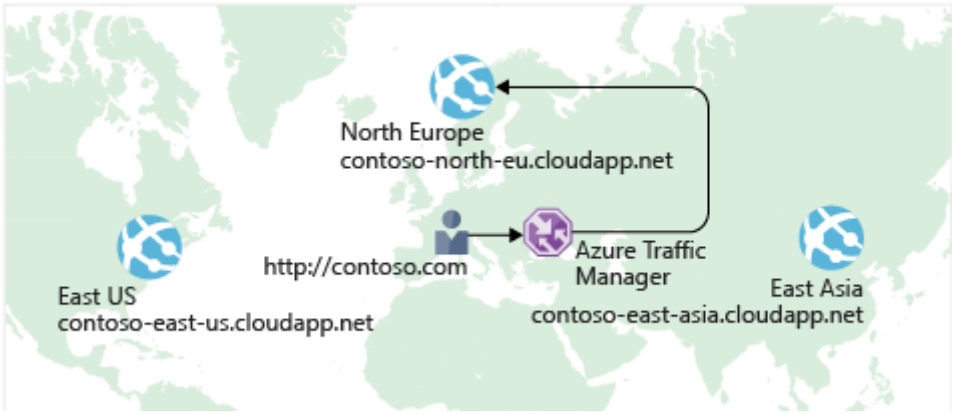
One way to reduce latency is to provide exact copies of your service in more than one region. The following illustration shows an example of global deployment.

The diagram shows your e-commerce site running in three Azure regions: East US, North Europe, and East Asia. Notice the DNS name for each. How can you connect users to the service that's closest geographically, but under the contoso.com domain?

Use Traffic Manager to route users to the closest endpoint

One answer is **Azure Traffic Manager**. Traffic Manager uses the DNS server that's closest to the user to direct user traffic to a globally distributed endpoint.

The following illustration shows the role of the Traffic Manager.



Traffic Manager doesn't see the traffic that's passed between the client and server. Rather, it directs the client web browser to a preferred endpoint. Traffic Manager can route traffic in a few different ways, such as to the endpoint with the lowest latency.

Although not shown here, this setup could also include your on-premises deployment running in California. You can connect Traffic Manager to your own on-premises networks, enabling you to maintain your existing data center investments. Or you can move your application entirely to the cloud. The choice is yours.

Compare Load Balancer to Traffic Manager

Azure Load Balancer distributes traffic within the same region to make your services more highly available and resilient. Traffic Manager works at the DNS level, and directs the client to a preferred endpoint. This endpoint can be to the region that's closest to your user.

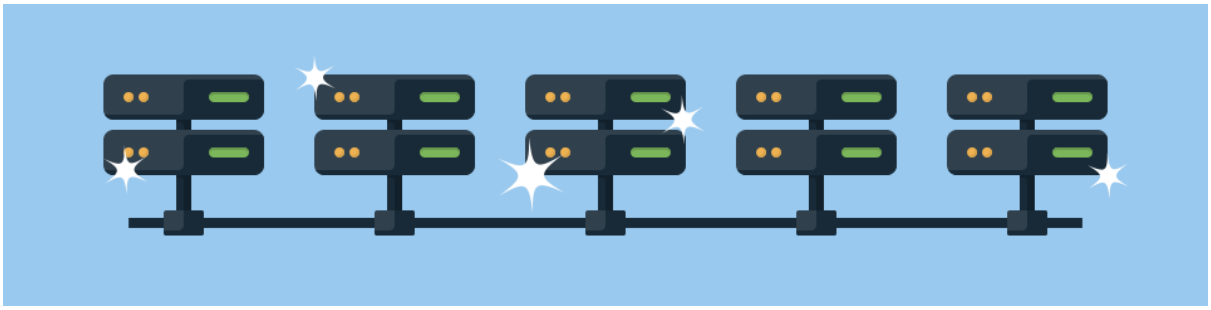
Load Balancer and Traffic Manager both help make your services more resilient, but in slightly different ways. When Load Balancer detects an unresponsive VM, it directs traffic to other VMs in the pool. Traffic Manager monitors the health of your endpoints. In contrast, when Traffic Manager finds an unresponsive endpoint, it directs traffic to the next closest endpoint that is responsive.

Summary

Geographic distance is one of the biggest factors that contributes to latency. With Traffic Manager in place, you can host exact copies of your service in multiple geographic regions. That way, users in the United States, Europe, and Asia will all have a good experience using your e-commerce site.

Summary

You learned just a few ways Azure networking can help reduce latency and make your apps and services more highly available.



With load balancing and global distribution in place, your e-commerce site is ready for the world. Users reach the domain that's closest geographically. Each domain has failover built in, helping every user have a great experience. The numbers are already showing increased traffic, and business is booming.