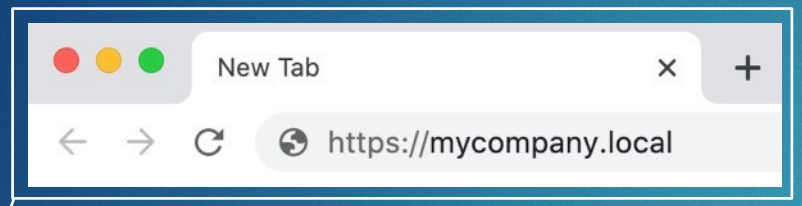
DNS, ELB and Auto Scaling

DNS and Amazon Route 53

The Domain Name System (DNS)

User enters **website address** in browser



Name	Type	Value
mycompany.local	A	192.168.0.1
emailserver.local	A	192.168.0.2

DNS Server

Domain name is resolved to the **IP address** of the webserver

Computer connects to **192.168.0.1**

Web Server

Amazon Route 53

Amazon Route 53

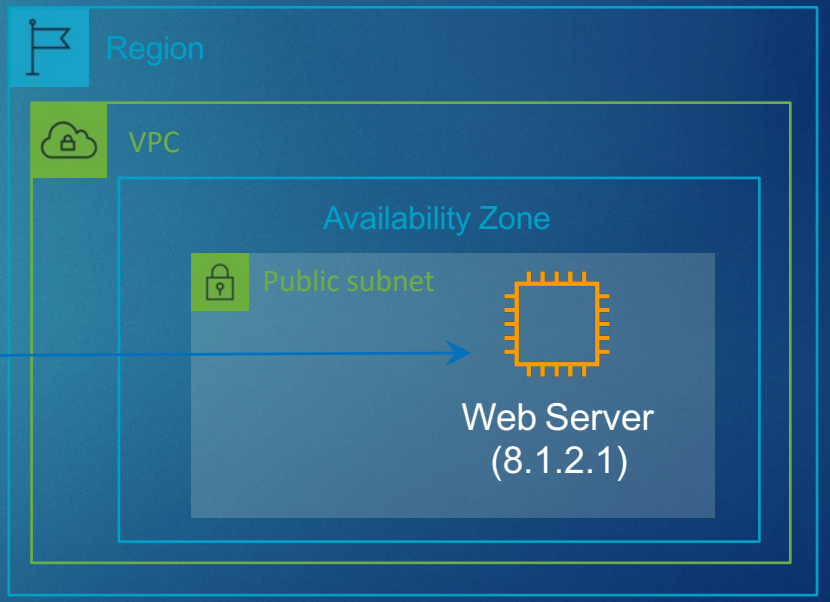
example.com

A **hosted zone** represents a set of records belonging to a domain

What's the IP address for **example.com**?

Address is 8.1.2.1

HTTP GET to **IP address** (8.1.2.1)



Amazon Route 53 Routing Policies

Routing Policy	What it does
Simple	Simple DNS response providing the IP address associated with a name
Failover	If primary is down (based on health checks), routes to secondary destination
Geolocation	Uses geographic location you're in (e.g. Europe) to route you to the closest region
Geoproximity	Routes you to the closest region within a geographic area
Latency	Directs you based on the lowest latency route to resources
Multivalue answer	Returns several IP addresses and functions as a basic load balancer
Weighted	Uses the relative weights assigned to resources to determine which to route to

Amazon Route Features

Amazon Route 53

Domain Registration

.net
.com
.org

Hosted zone

example.com
dctlabs.com

Health Checks

EC2 Instances

Traffic Flow

Register Domain with Route 53 (Optional)

Elasticity: Scaling Up vs Out

Scaling Up (vertical scaling)

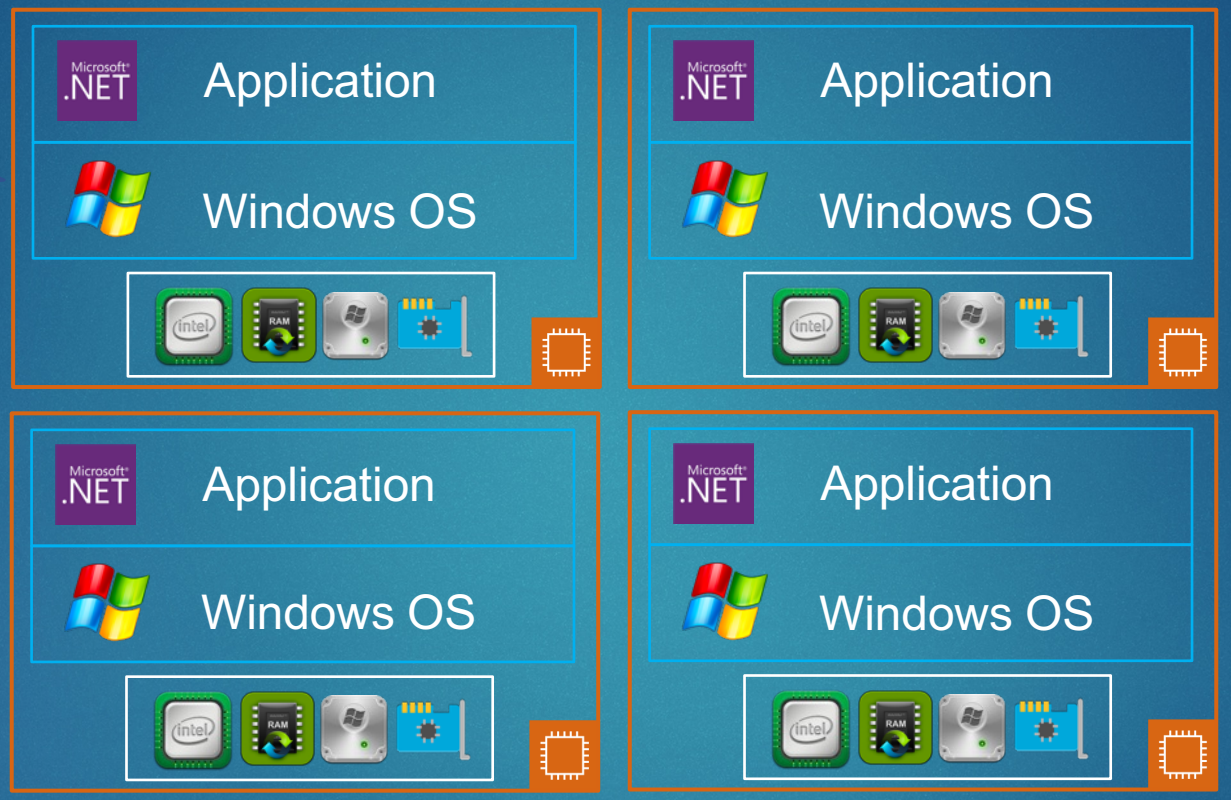
Scaling Up (vertical scaling)

Scaling up means **adding** resources to the instance

Limitation is that you have a **single point of failure** (SPOF)

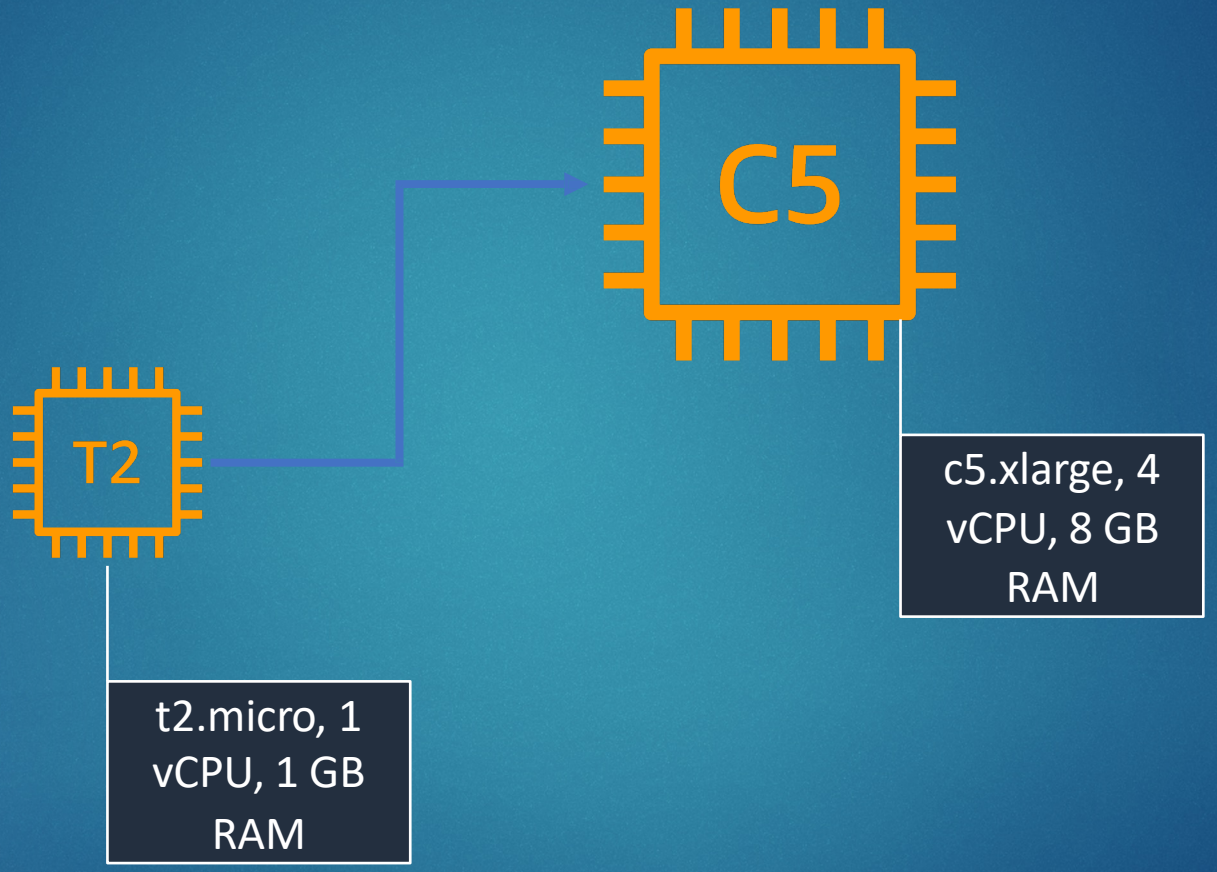
Scaling Out (horizontal scaling)

Scaling out provides greater **resiliency**

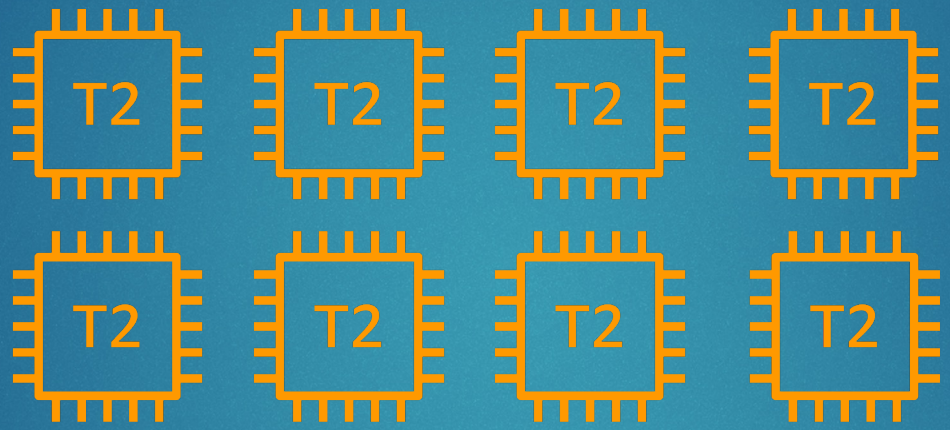


Scaling out can be used to add almost unlimited capacity

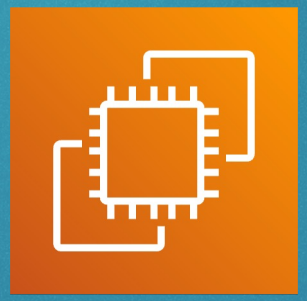
Scaling Up (vertical scaling)



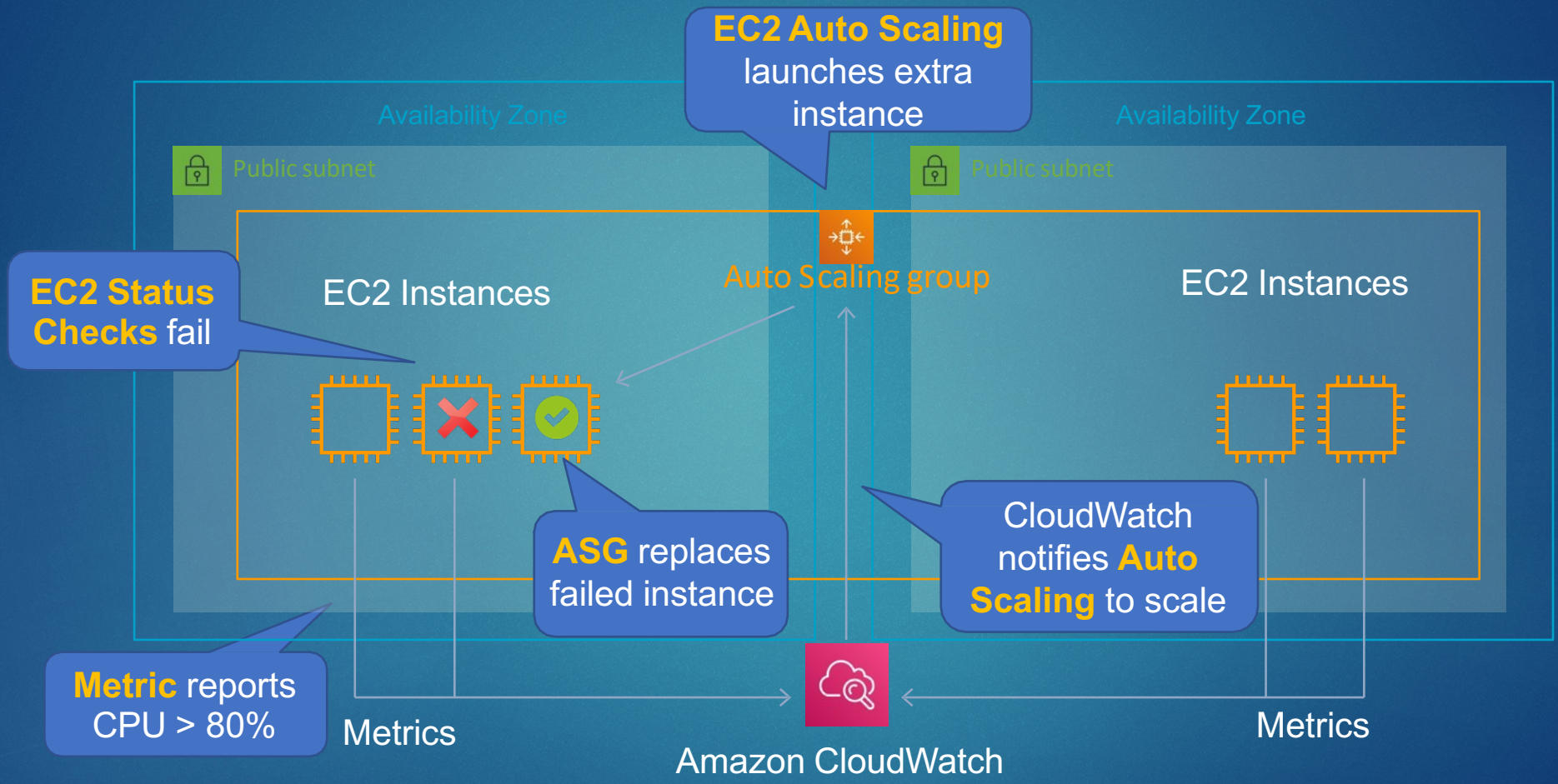
Scaling Out (horizontal scaling)



Amazon EC2 Auto Scaling



Amazon EC2 Auto Scaling

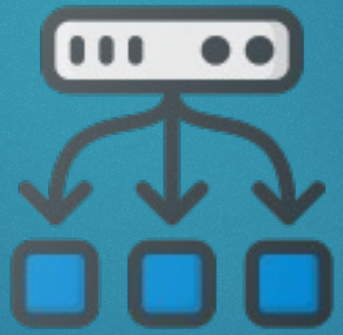


Amazon EC2 Auto Scaling

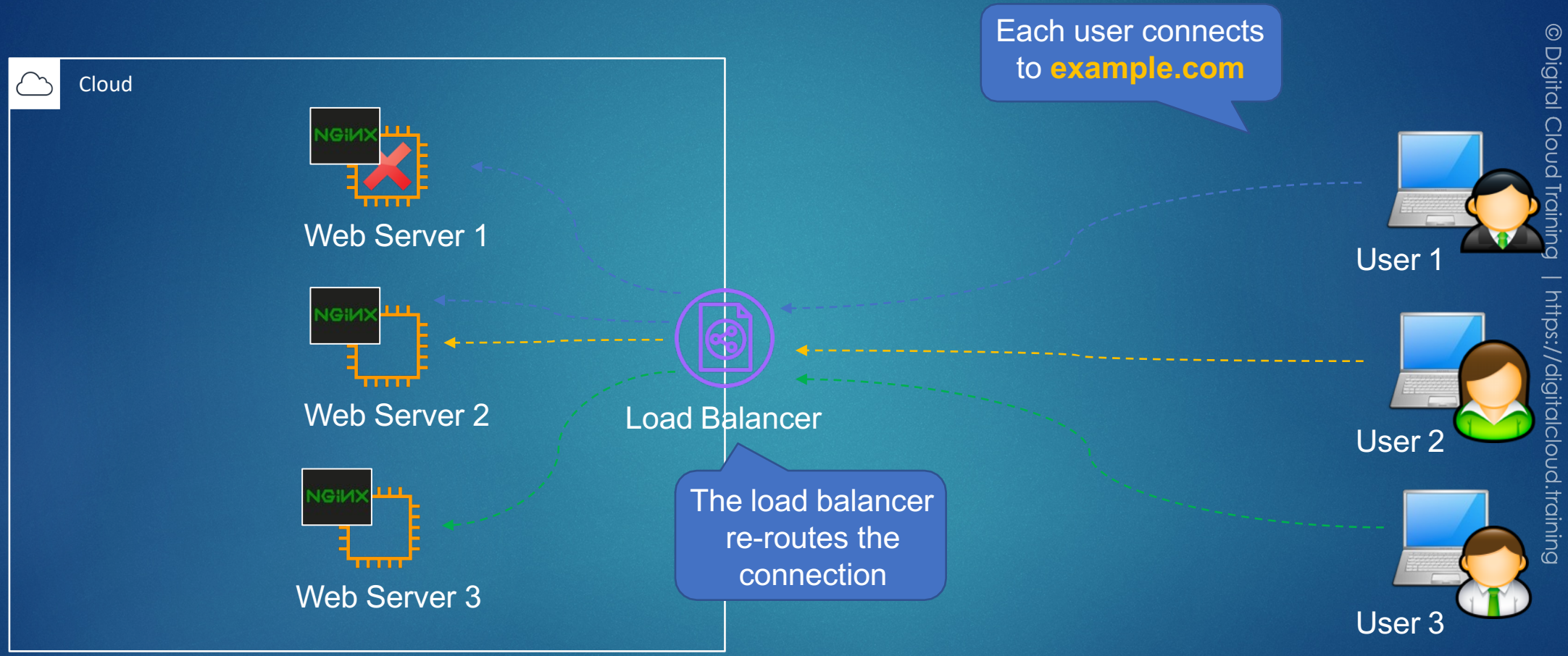
- EC2 Auto Scaling **launches** and **terminates** instances dynamically
- Scaling is horizontal (scales out)
- Provides **elasticity** and **scalability**
- Responds to EC2 status checks and CloudWatch metrics
- Can scale based on demand (performance) or on a schedule
- Scaling policies define how to respond to changes in demand

Create an Auto Scaling Group

Load Balancing and High Availability



Load Balancing and High Availability

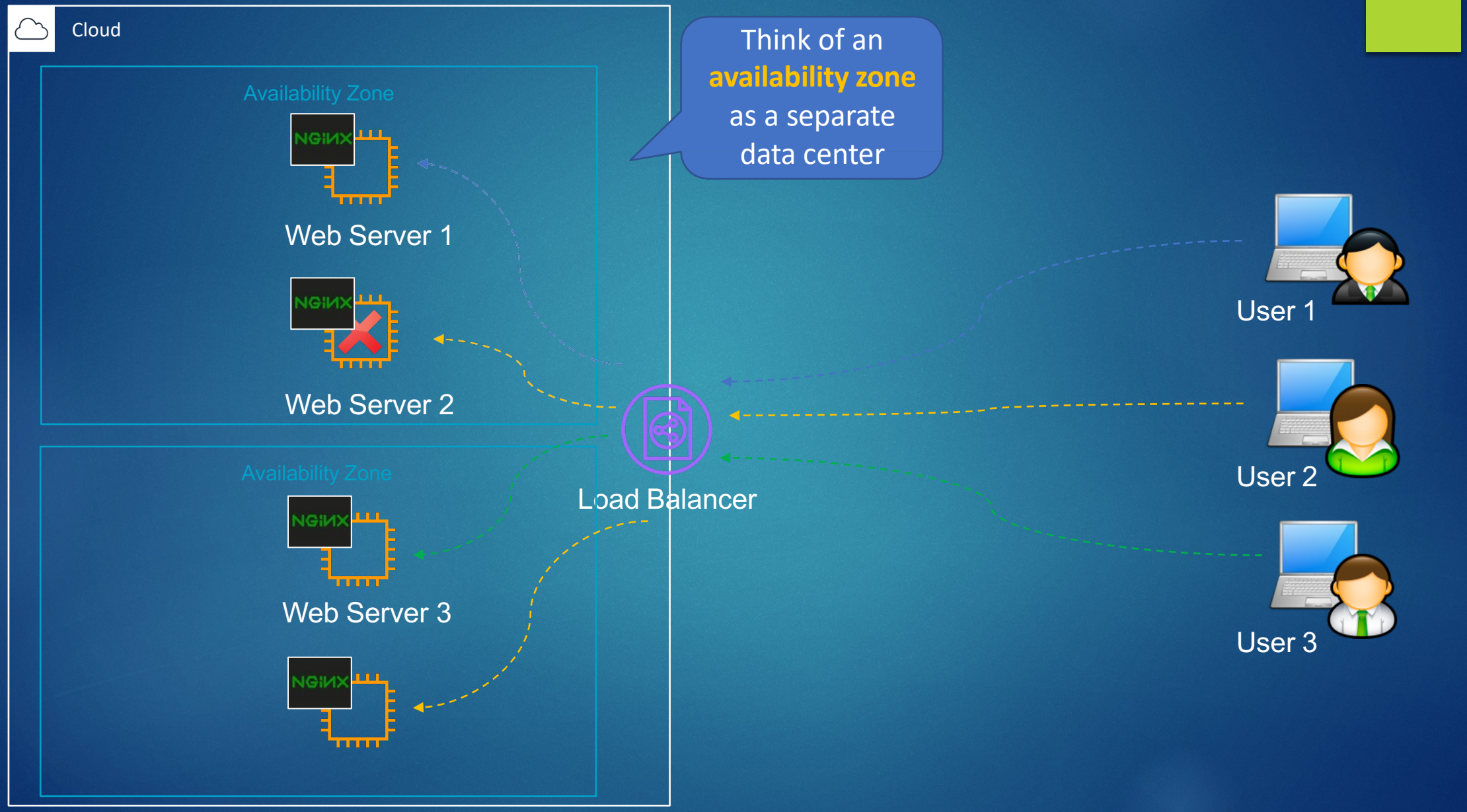


Fault Tolerance

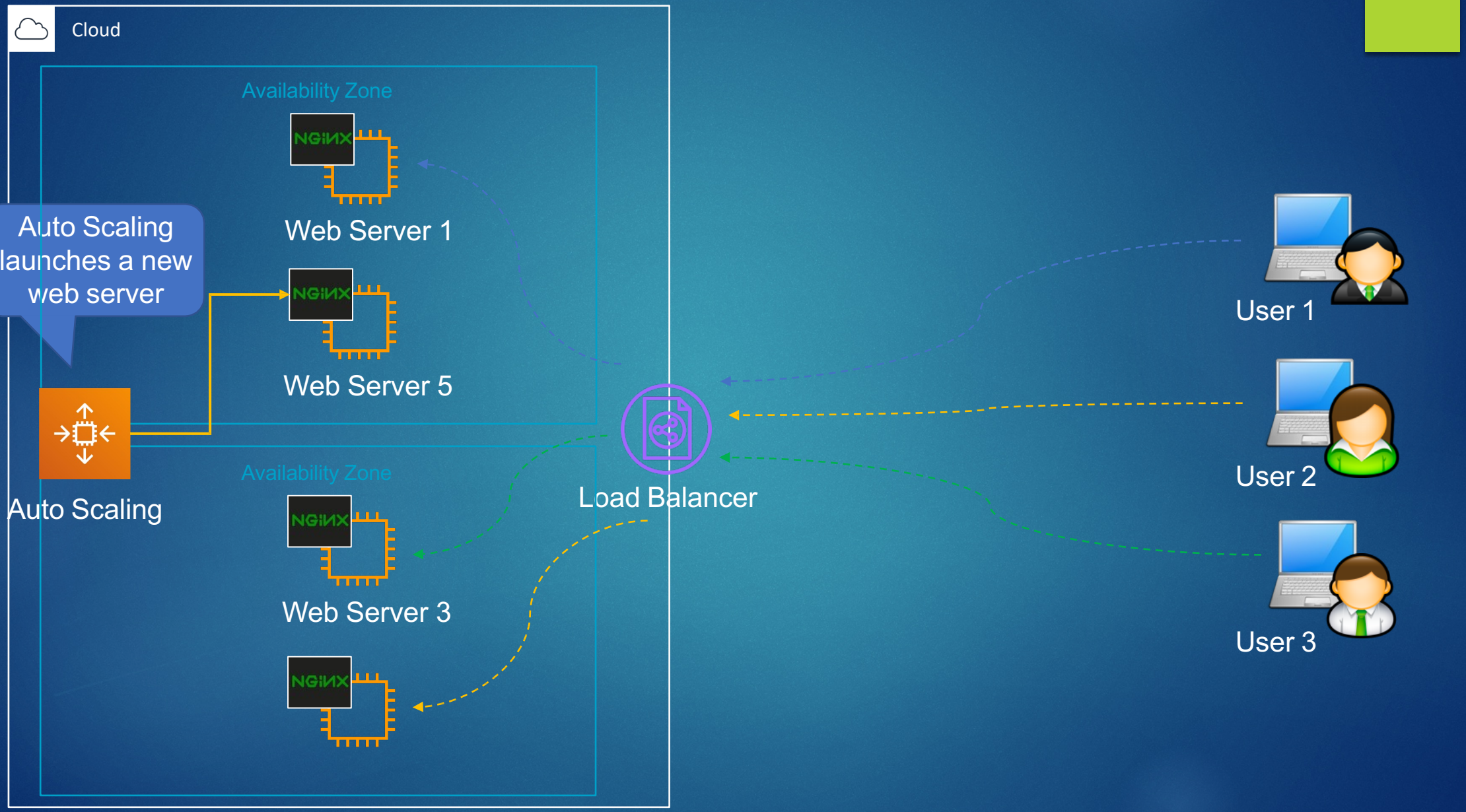
Redundant components allow the system to continue to operate

The system may fail if there is no built-in **redundancy**

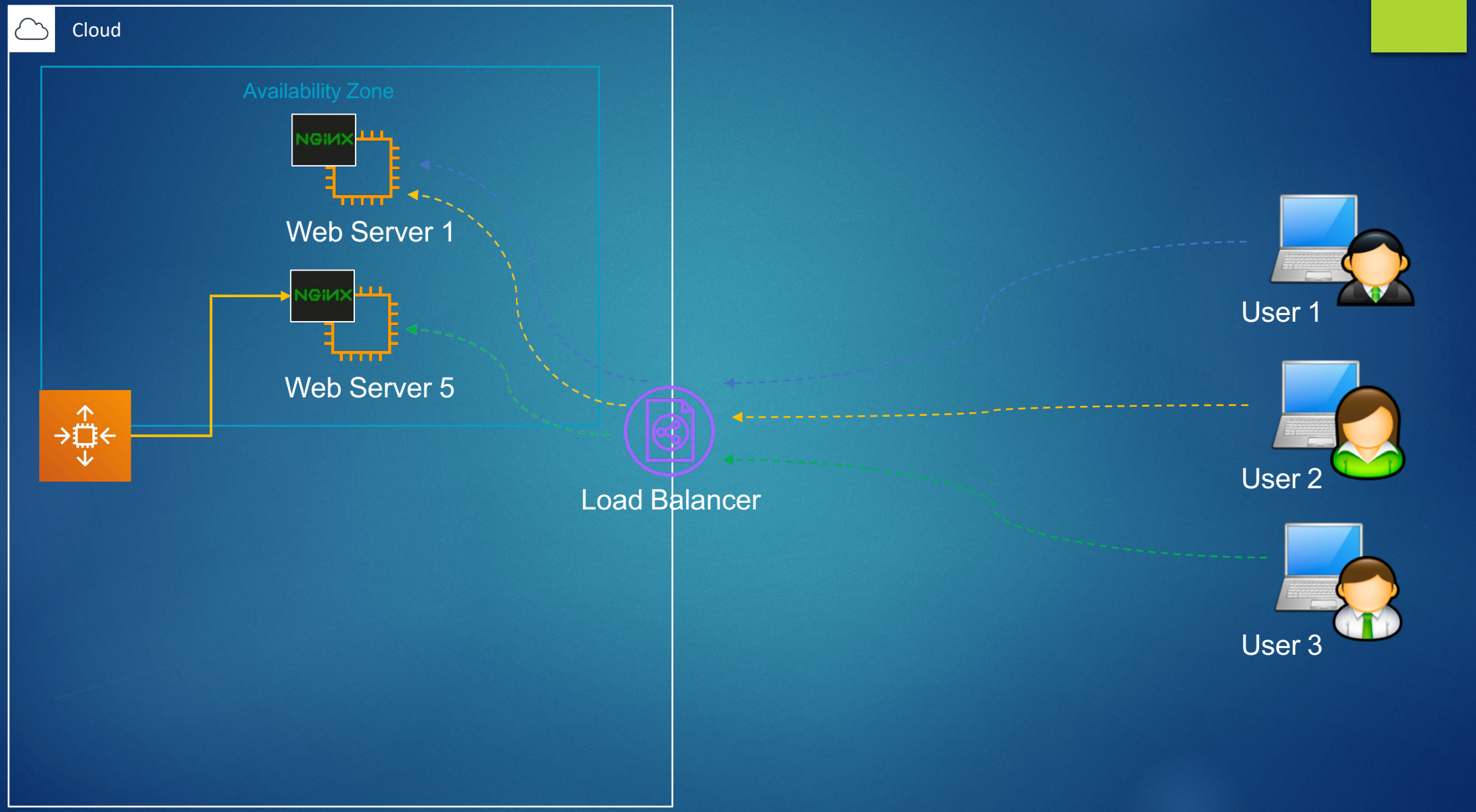
High Availability and Fault Tolerance



High Availability and Fault Tolerance

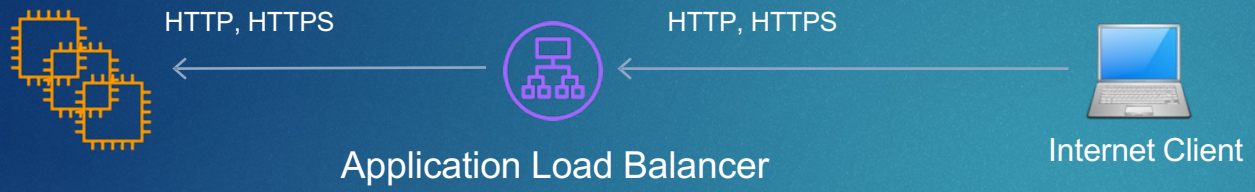


High Availability and Fault Tolerance



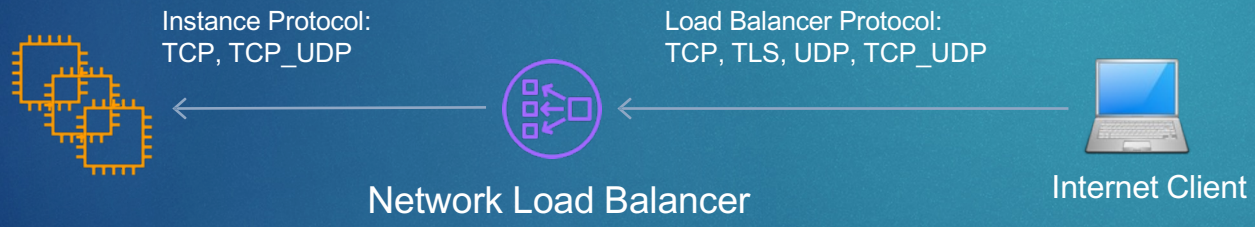
Amazon Elastic Load Balancer (ELB)

Types of Elastic Load Balancer (ELB)



Application Load Balancer

- Operates at the request level
- Routes based on the content of the request (layer 7)
- Supports advanced routing

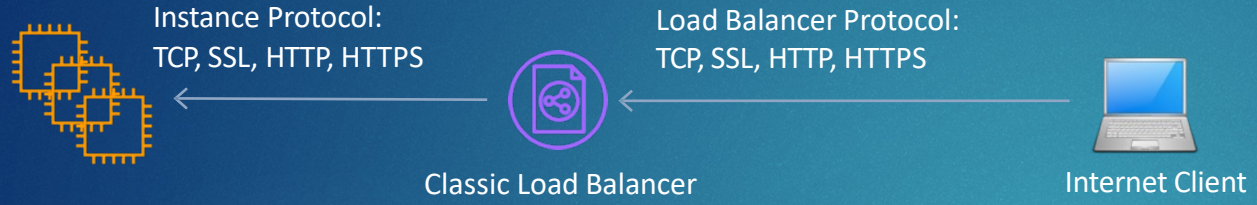


Network Load Balancer

- Operates at the connection level
- Routes connections based on IP protocol data (layer 4)
- Offers ultra high performance, low latency and TLS offloading at scale

Types of Elastic Load Balancer (ELB)

Old and **shouldn't** be the exam anymore



Classic Load Balancer

- Old generation; not recommended for new applications
- Performs routing at Layer 4 and Layer 7
- Use for existing applications running in EC2-Classic

Gateway Load Balancer

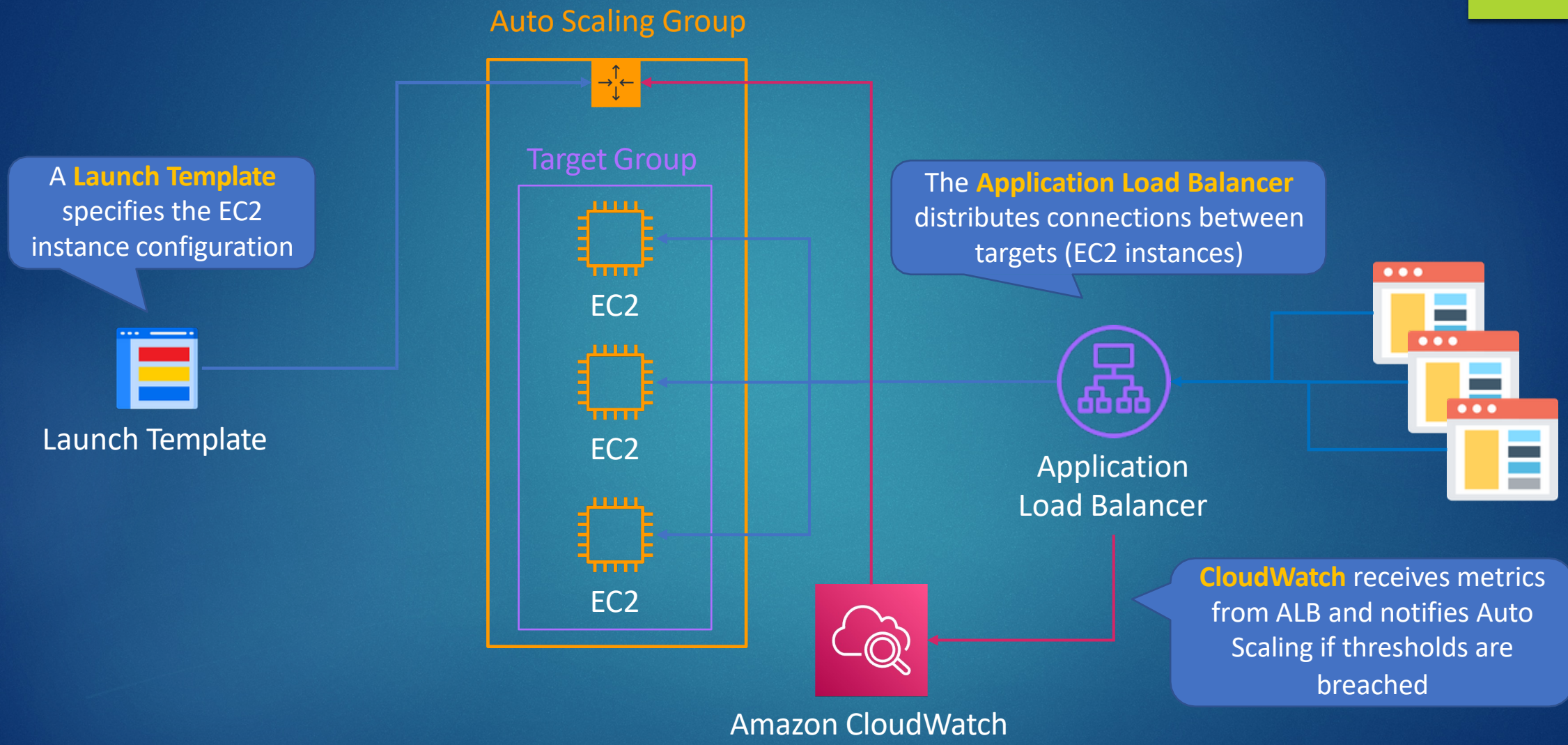
- Used in front of virtual appliances such as firewalls, IDS/IPS, and deep packet inspection systems

New and **not** yet on the exam

Attach an Application Load Balancer

Elastically Scale the Application

Elastically Scale the Application



Scaling Policies

Scaling Policies

- **Target Tracking** – Attempts to keep the group at or close to the metric
- **Simple Scaling** – Adjust group size based on a metric
- **Step Scaling** – Adjust group size based on a metric – adjustments vary based on the size of the alarm breach
- **Scheduled Scaling** – Adjust the group size at a specific time