Scalability is the ability to add

or remove virtual machines, scaling out,

or increase the resources

on a single virtual machine, scaling up.

= = = = = = =

In cloud computing, scalability is the process

of adding more resources on an as-needed basis.

This is important because, in a lot of cases,

you don't know when your users are using the system.

At various times, you can have an influx of users,

which suddenly means that the current servers,

being virtual machines or other services,

will run out of resources to serve all those requests.

What would happen with traditional,

self-hosted hardware is that your servers

would either deny the requests

or just fall over (gags).

Not ideal.

So because Azure has a perceivable

infinite number of resources, such as servers,

when more traffic arrives to your application,

Azure can add more resources.

This is scalability.

Azure can scale the user resources

with the needs of your application,

service, or product.

There are, of course, different ways

that you can scale resources.

Scaling out is adding

more resources to the pool of available resources

for your application.

like adding all identical,

but they get added real quick.

When the resource need goes down,

the resources are removed again.

Of course, you can also scale up,

and that is making current resources bigger.

You can, of course, also scale down.

That is, getting a smaller resource.

Autoscaling is when either approach happens automatically,

according to rules that you've set.

This could be that every day at 9 a.m.,

you scale up,

or, if your CPU usage exceeds 70%,

you scale out and get more resources.

That's scalability.

Scalability is scaling out, up, or down,

and doing so automatically to be completely capable

of handling increased or decreased demand.

It is one of the terms for describing

how responsive Azure is to change in demand.