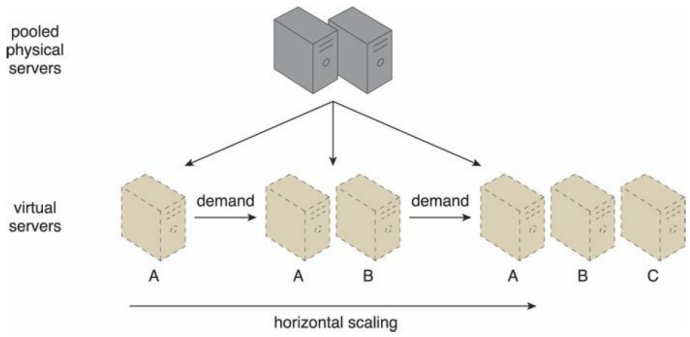
Hongfei Xie & Pouria Ghadimi

References: Cloud Computing

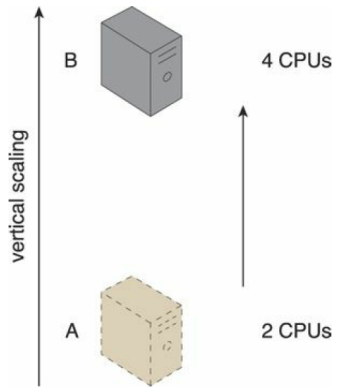
Horizontal Scaling (scale by adding more machines)

The allocating or releasing of IT resources that are of the same type is referred to as horizontal scaling. The horizontal allocation of resources is referred to as scaling out and the horizontal releasing of resources is referred to as scaling in. Horizontal scaling is a common form of scaling within cloud environments.



Vertical Scaling (you scale by adding more power (CPU, RAM) to an existing machine)

When an existing IT resource is replaced by another with higher or lower capacity, vertical scaling is considered to have occurred. Specifically, the replacing of an IT resource with another that has a higher capacity is referred to as scaling up and the replacing an IT resource with another that has a lower capacity is considered scaling down. Vertical scaling is less common in cloud environments due to the downtime required while the replacement is taking place.



Challenge:

Horizontal: more machines.

Vertical: power, cooling, and hardware options.

Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity