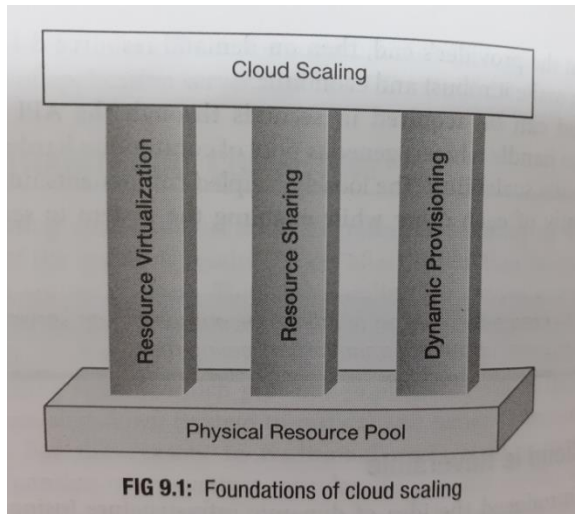
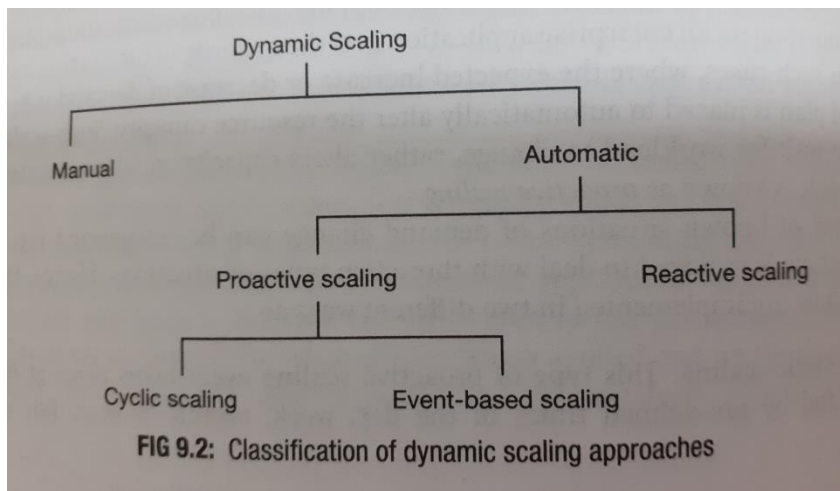


## Scaling: grow and shrink

### Foundation of Scaling:



### Strategies:



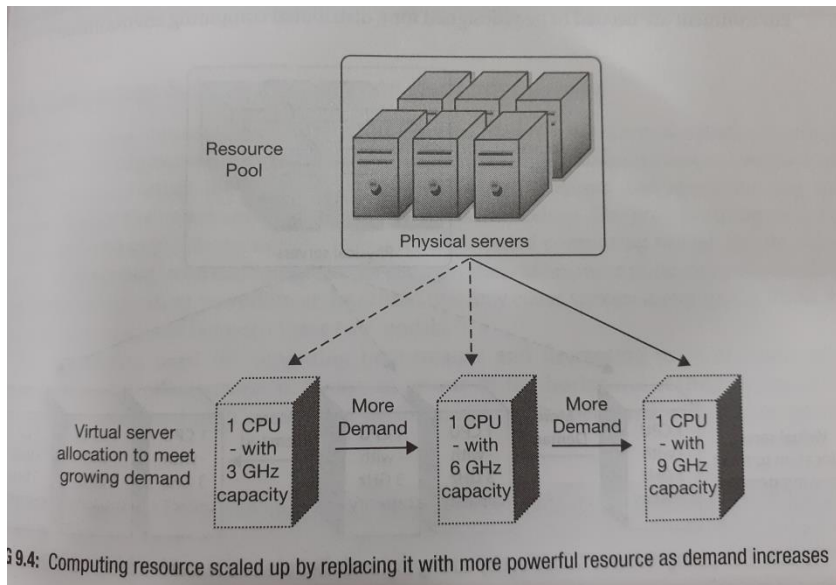
**Reactive Scaling**, or also called Auto Scaling simply monitors your applications and adjusts its capacity to maintain optimum performance at supposedly minimum cost

**Proactive Scaling** As it sounds from the name, proactive or scheduled scaling allows you to scale your application resources based on known load that will appear on future.

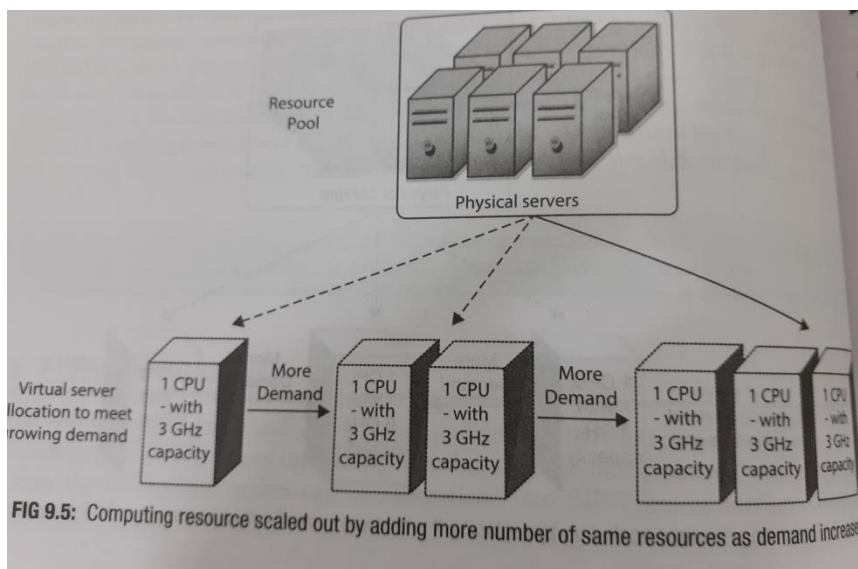
**Proactive Cycle Scaling:** automatically start up and shutdown vms during peak periods, e.g. 9am-5pm mon-fri.

**Proactive Event-Based Scaling:** Can automatically scale in anticipation peaks caused by certain events. E.g. black-Friday, boxing day, half-price sale days.

## Vertical scaling or Scaling Up



## Horizontal scaling or Scaling Out



<b>Vertical Scaling</b>	<b>Horizontal Scaling</b>
It is known as Scaling up.	It is known as Scaling out.
This approach involves replacement of components or resource nodes.	This approach involves introduction of additional components or resource nodes.
It can be implemented in any type of computing environment.	It can only be implemented in distributed computing environment.
It has less management complexity.	The management of larger numbers of nodes increases system complexity.
It may cause service interruption as system needs to restart after replacing component.	It does not cause service interruption. No system restart is required.
It has less influence on application architecture.	This approach has a more fundamental influence on application architecture.
A vertically scalable application tends to run on high-end hardware.	A horizontally scalable application tends to run on low-end (commodity) hardware.
Here, the loads are concentrated.	It spreads the load.
It requires the specialized hardware components.	It can be done using normal commodity hardware components.
Upgrading the capacity of resource beyond a certain level can become very expensive in this type of scaling.	It is always less expensive and the expense is always proportional to resource capacity.
The expansion is limited by any resource's maximum capacity.	The expansion does not depend on available capacity of hardware components.
It is not a long term solution for scaling.	It provides a long term solution for scaling.

### Choice Vertical or Horizontal?

#### References

- *Cloud Computing*, Sandeep Bhowmik