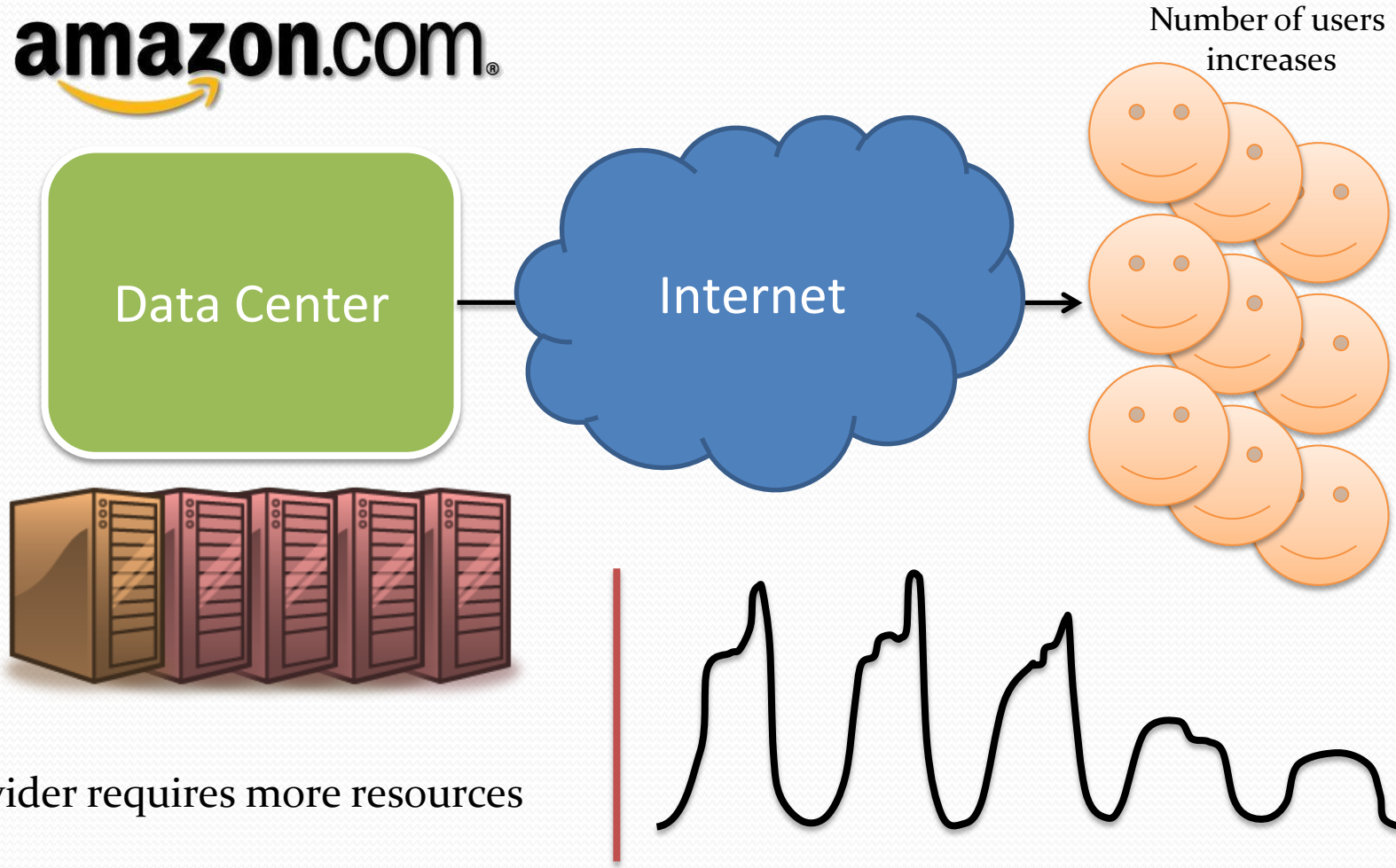
بسمه تعالی

مدیریت منابع و مقیاس پذیری در رایانش ابری

ارائه دهنده : محمد فرجی مهماندار

WWW.CFSCO.IR

To meet the fluctuating application workload demands, dynamic provisioning is essential.



Resource Management Approaches in Cloud Computing

- Cloud computing is being viewed as the technology of today and the future.
- This technology allows for provisioning of various resources such as virtual machines (VM), physical machines, processors, memory, network, storage and software as per the needs of customers.
- Application providers (AP), who are customers of the CP, deploy applications on the cloud infrastructure and then these applications are used by the end-users.
- To meet the fluctuating application workload demands, dynamic provisioning is essential.
- Cloud dynamic provisioning is explained by considering resources, stakeholders, techniques, technologies, algorithms, problems, goals and more.

CLASSIFYING CLOUD PROVISIONING

1) VM provisioning:

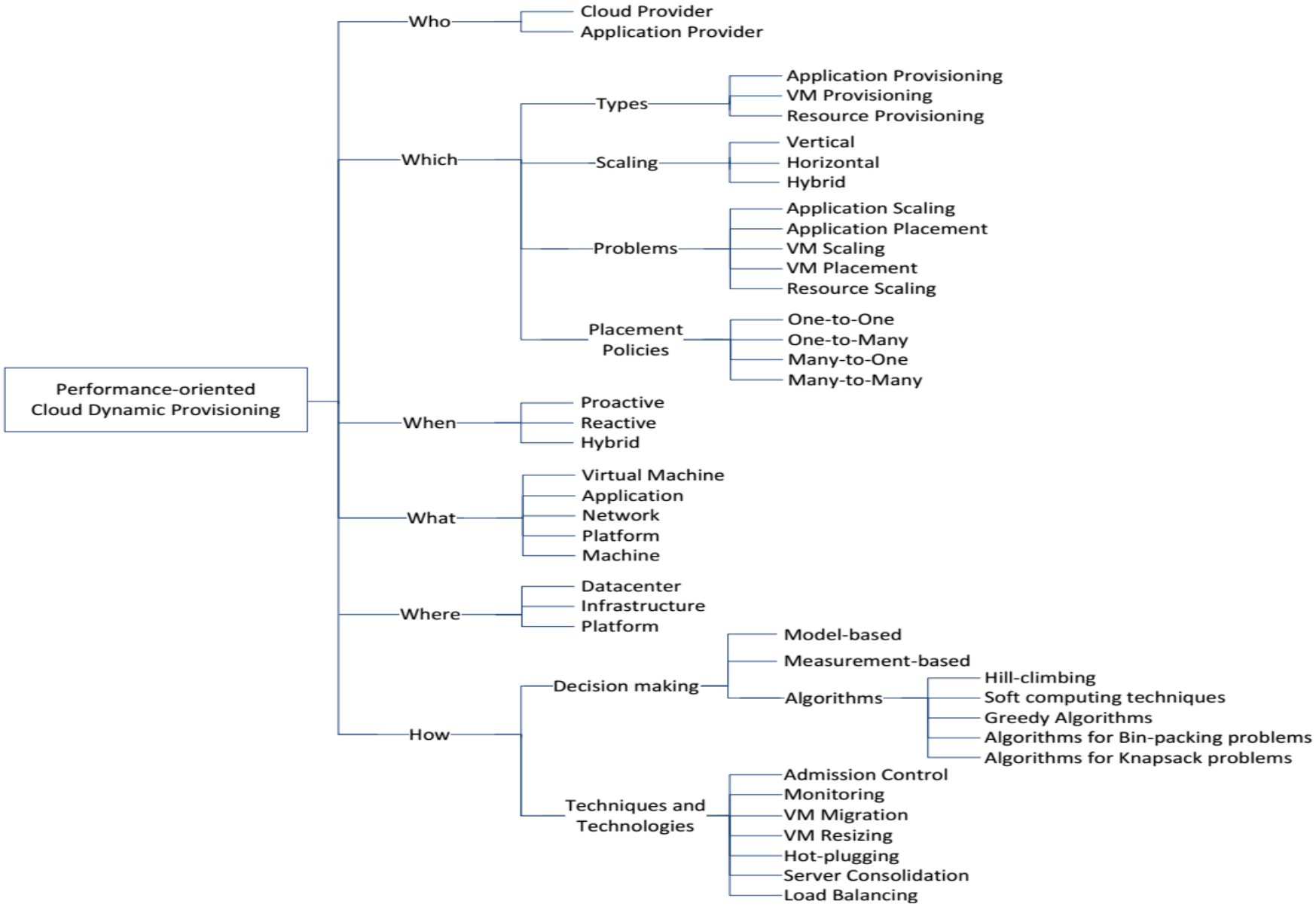
creation of VMs to meet software and hardware based requirements of an application such that given performance levels are achieved.

2) Resource provisioning:

association of created VMs to adequate hardware resources.

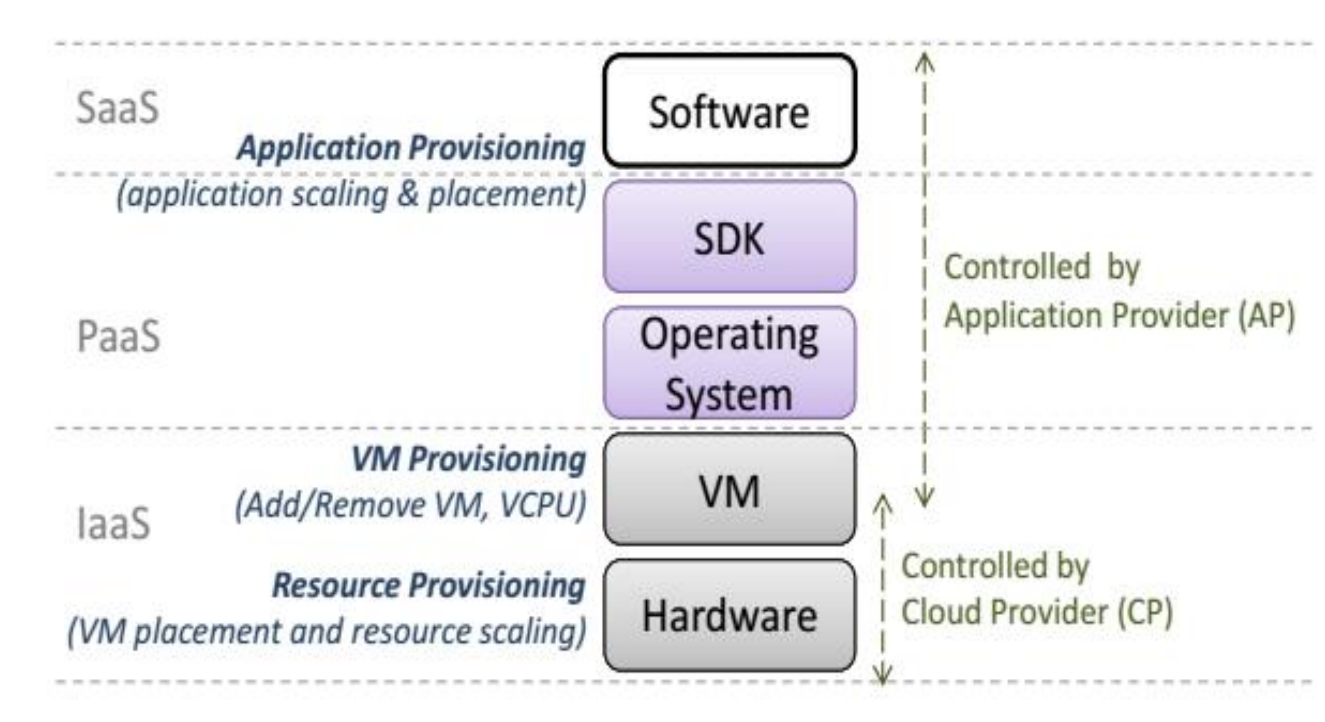
3) Application provisioning:

application deployment in the VMs and the subsequent association of the requests received to those applications.


Performance-oriented Cloud Dynamic Provisioning
Who
Cloud Provider
Application Provider
Which
Types
Application Provisioning
VM Provisioning
Resource Provisioning
Scaling
Vertical
Horizontal
Hybrid
Problems
Application Scaling
Application Placement
VM Scaling
VM Placement
Resource Scaling
Placement Policies
One-to-One
One-to-Many
Many-to-One
Many-to-Many
When
Proactive
Reactive
Hybrid
What
Virtual Machine
Application
Network
Platform
Machine
Where
Datacenter
Infrastructure
Platform
How
Decision making
Model-based
Measurement-based
Algorithms
Hill-climbing
Soft computing techniques
Greedy Algorithms
Algorithms for Bin-packing problems
Algorithms for Knapsack problems
Techniques and Technologies
Admission Control
Monitoring
VM Migration
VM Resizing
Hot-plugging
Server Consolidation
Load Balancing

Who performs the provisioning (the decision making entity).

These would be entities that make decisions about when resources need to be added and removed. AP and CP are two such entities.



Which provisioning types, scaling, problems and policies exist.

In simple classification, provisioning may be classified as Static Resource Provisioning and Dynamic Resource Provisioning.

❑In another classification, provisioning may be classified as application, VM and resource provisioning: Related to these provisioning types are problems such as application scaling, application placement, VM scaling, VM placement and resource scaling.

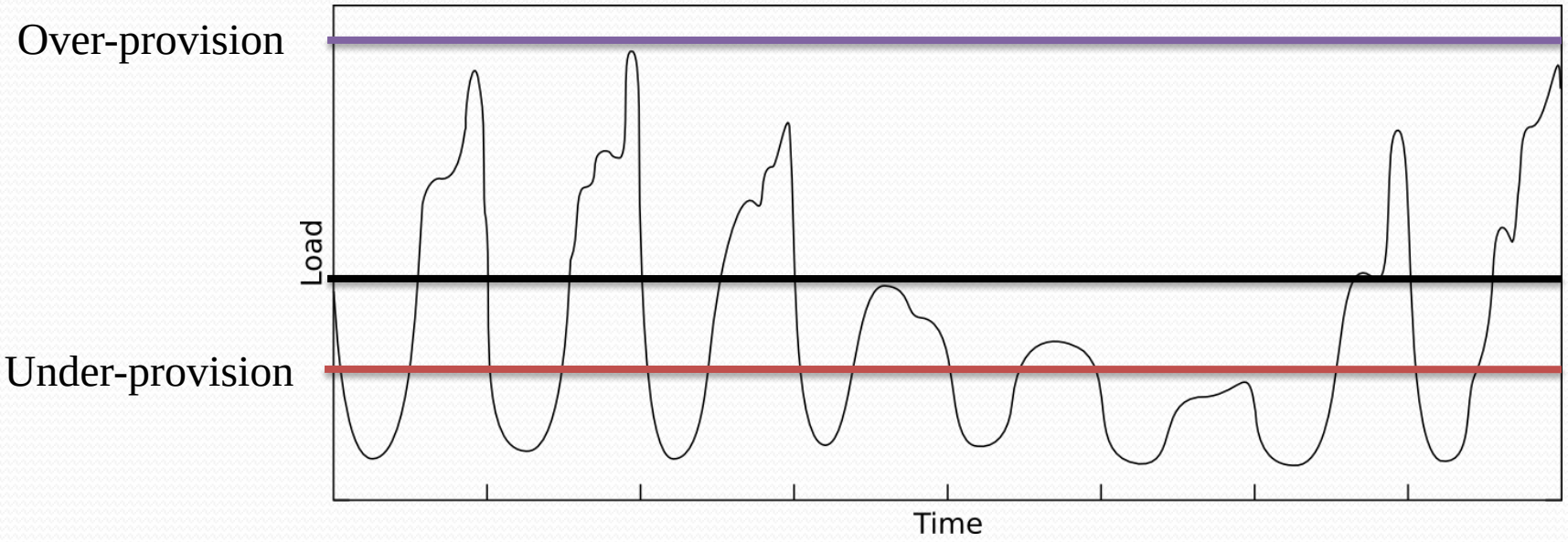
❑ Furthermore, resource scaling may be classified either as horizontal or vertical : where the former type of scaling is based on adding or removing of new or existing resources (e.g. VMs) and the latter is associated with adding of resources to the existing live VMs and machines.

❑ The relationships between resources and placement policies :

One-to-one, one-to-many, many-to-one, and many-to-many policies could be used to define different placements of VMs on physical machines and of applications on VMs.

Provisioning types : Static Provisioning Vs Dynamic Provisioning

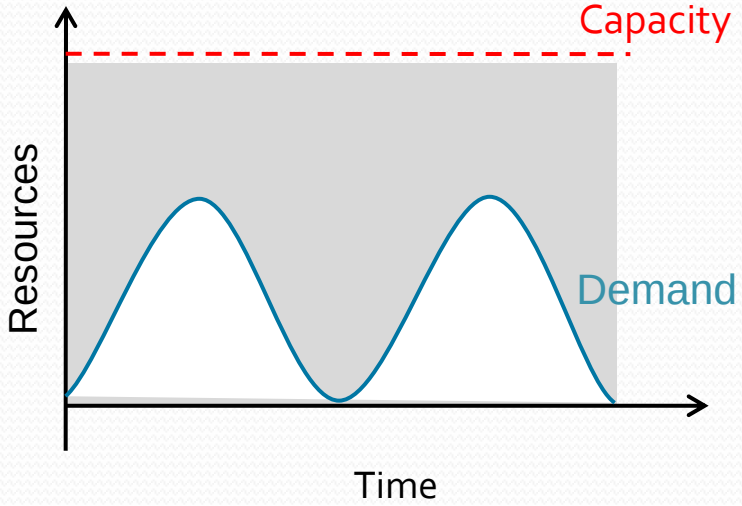
- **Challenges:**
  - **Over-Provisioning**
  - **Under-Provisioning**
  - **Oscillation**



Static Provisioning

اضافه تامین(Over-Provisioning) یا (Under-Utilization)

- رزرو کردن حداکثری منابع به اندازه زمان اوج تقاضا
  - هدر رفتن منابع(کاهش بهره وری)
  - کاهش تعداد تخطی از SLA
  - افزایش هزینه برای خرید منابع (کاهش سود)



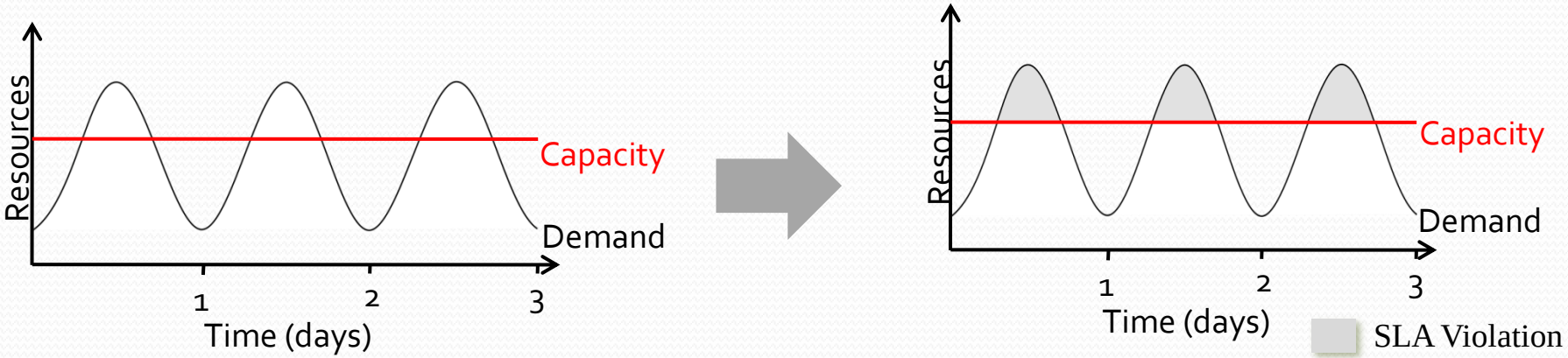
Unused resources

Over-Provisioning={Utilization↓, SLA-Violation↓, Profit↓}

Static Provisioning

کسر تامین(Under-Provisioning) یا (Full-Utilization)

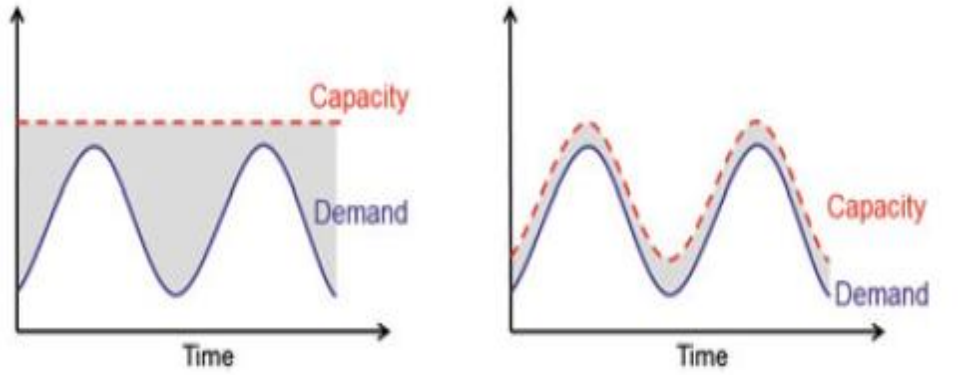
- رزرو کردن منابع به اندازه کافی (متوسط بار کاری)
  - بهره وری کامل از منابع
  - افزایش تعداد تخطی SLA (کاهش شهرت)
  - رد پذیرش مشتریان جدید و از دست رفتن مشتریان(کاهش سود)

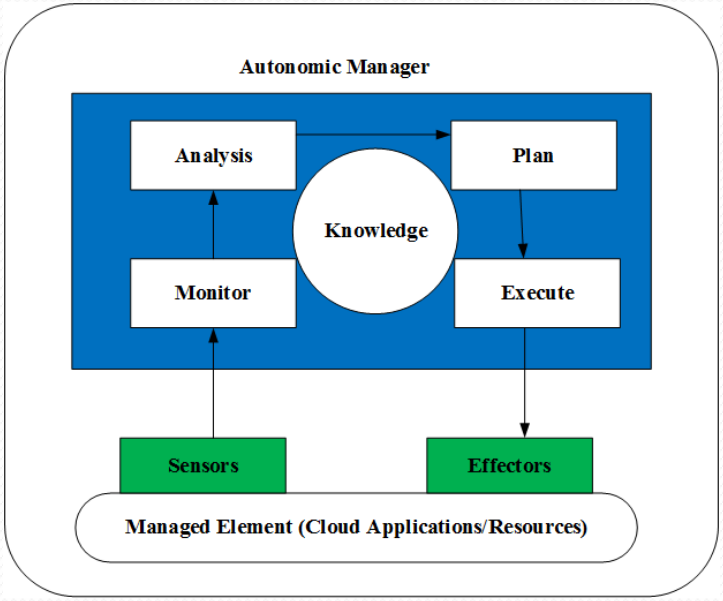


Under-Provisioning={Utilization↑,SLA-Violation↑,Profit↓}

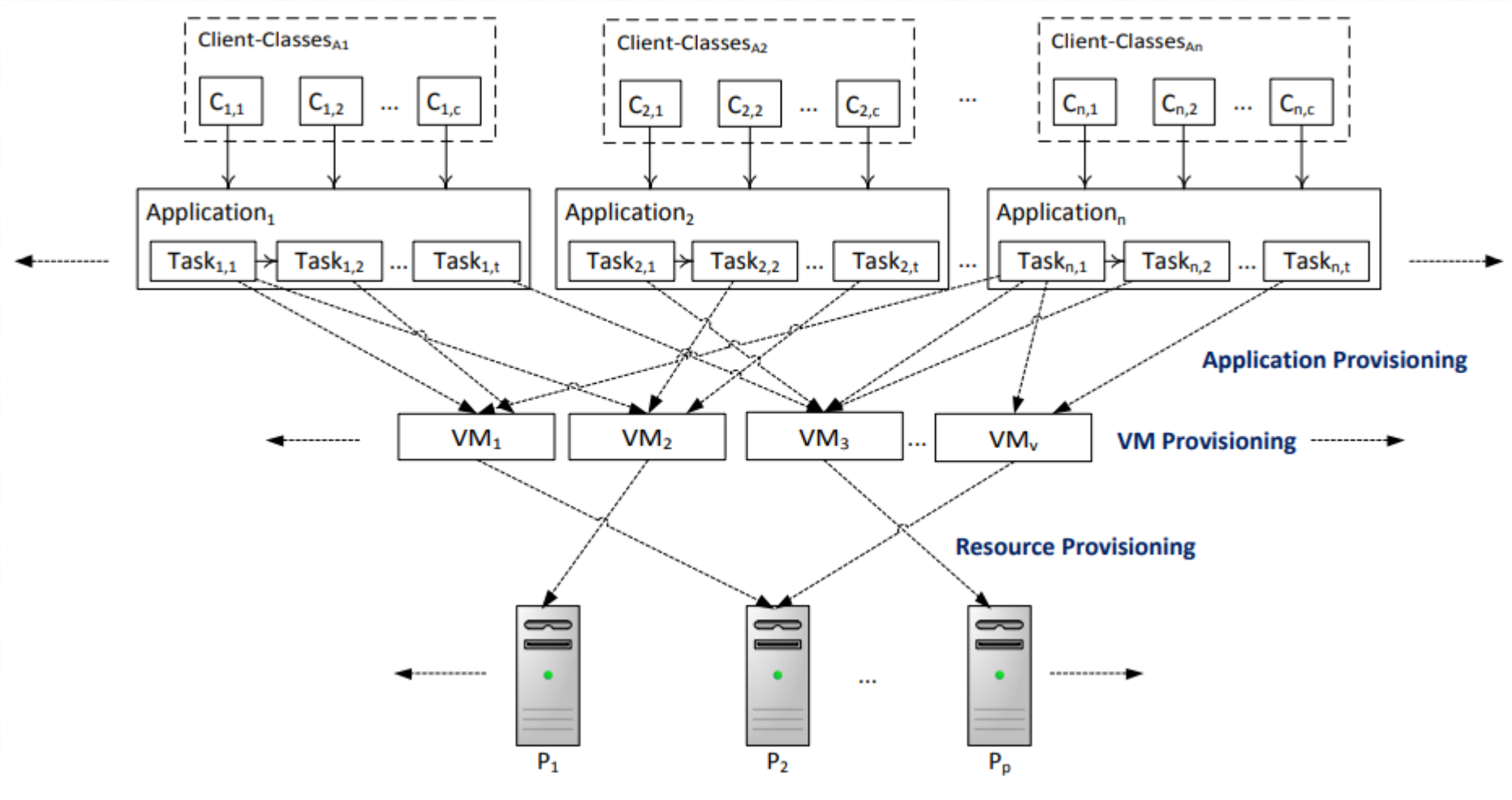
Dynamic Provisioning

- In contrast, dynamic provisioning solves the under and over provisioning of resources by adjusting resource allocations to the changing workload that the system receives.
- dynamically adding and removing cloud resources to handle the fluctuating Internet user request demands of the applications





Placement Policies



Provisioning Problems

dynamic provisioning to be associated with five problems:

1) **Application scaling (AppScale):** determining the increase and decrease in the number of applications or units that applications are composed of (e.g. by addition of replicas of processes or application tiers).
2) **Application placement (AppPlace):** determining the placement of application units on VMs or on physical machines; the latter is for when VMs are not employed.
3) **VM Scaling (VmScale):** determining the increase and decrease in the number of VMs and associated resources such as virtual CPUs, memory etc.
4) **VM placement (VmPlace):** determining the placement of VMs on the machines, i.e. the allocation of resources to the VMs.
5) **Resource Scaling (ResScale):** determining the increase and decrease in the number of operating resources, which the applications and/or the VMs will run on and utilize.

Dynamic provisioning and associated Problems

Provisioning	Associated Problems	Responsibility
Application Provisioning	*AppScale* AND/OR *AppPlace*	AP
VM Provisioning	*VmScale*	AP
Resource Provisioning	*VmPlace* AND/OR *ResScale*	CP

Placement relationships and notations

Notation	Relationship	Description
$A \xrightarrow{1..1} B$	*one-to-one*	each A may be placed on one-and-only-one B
$C \xrightarrow{1..\infty} D$	*one-to-many*	each C may be placed on multiple D's, but multiple C's shall not reside within each D
$E \xrightarrow{\infty..1} F$	*many-to-one*	multiple E's may reside on one F, but these E's shall not span across multiple F's, except through replication of E
$G \xrightarrow{\infty..\infty} H$	*many-to-many*	multiple G's may reside on one H, and each G may span multiple H's

WHEN IS THE PROVISIONING DECISION MADE?

Proactive provisioning is made possible through use of workload predictors, performance models and system monitors. Through workload prediction, the future incoming workload can be predicted and this would feed as input to the performance model. The parameters of the performance model would be determined through monitoring the system and through prediction modules. The performance metrics thus obtained from solving the model would indicate if QoS objectives would be satisfied through the current system configuration.

In contrast to proactive provisioning,

Reactive provisioning proceeds with modification of system configuration as the workload changes.

WHAT RESOURCES ARE PROVISIONED?

Various resources are provided by clouds as services:

1) VMs and associated resources are provisioned through IaaS
2) Development platform is made available through PaaS
3) Software is accessible through SaaS

HOW ARE THE PROVISIONING PROBLEMS SOLVED?

- ❑Algorithms
- ❑Techniques and Technologies

What is Scalability?

Scalability is a term used to describe how the application will handle increased loads of traffic volume

Scalability vs. Elasticity

- The purpose of Elasticity is to match the resources allocated with actual amount of resources needed at any given point in time.
- Scalability handles the changing needs of an application within the confines of the infrastructure via statically adding or removing resources to meet applications demands if needed. In most cases, this is handled by adding resources to existing instances—called scaling up or vertical scaling—and/or adding more copies of existing instances—called scaling out or horizontal scaling.
- In addition, scalability can be more granular and targeted in nature than elasticity when it comes to sizing.

What Type of Scalability? Vertical vs. Horizontal

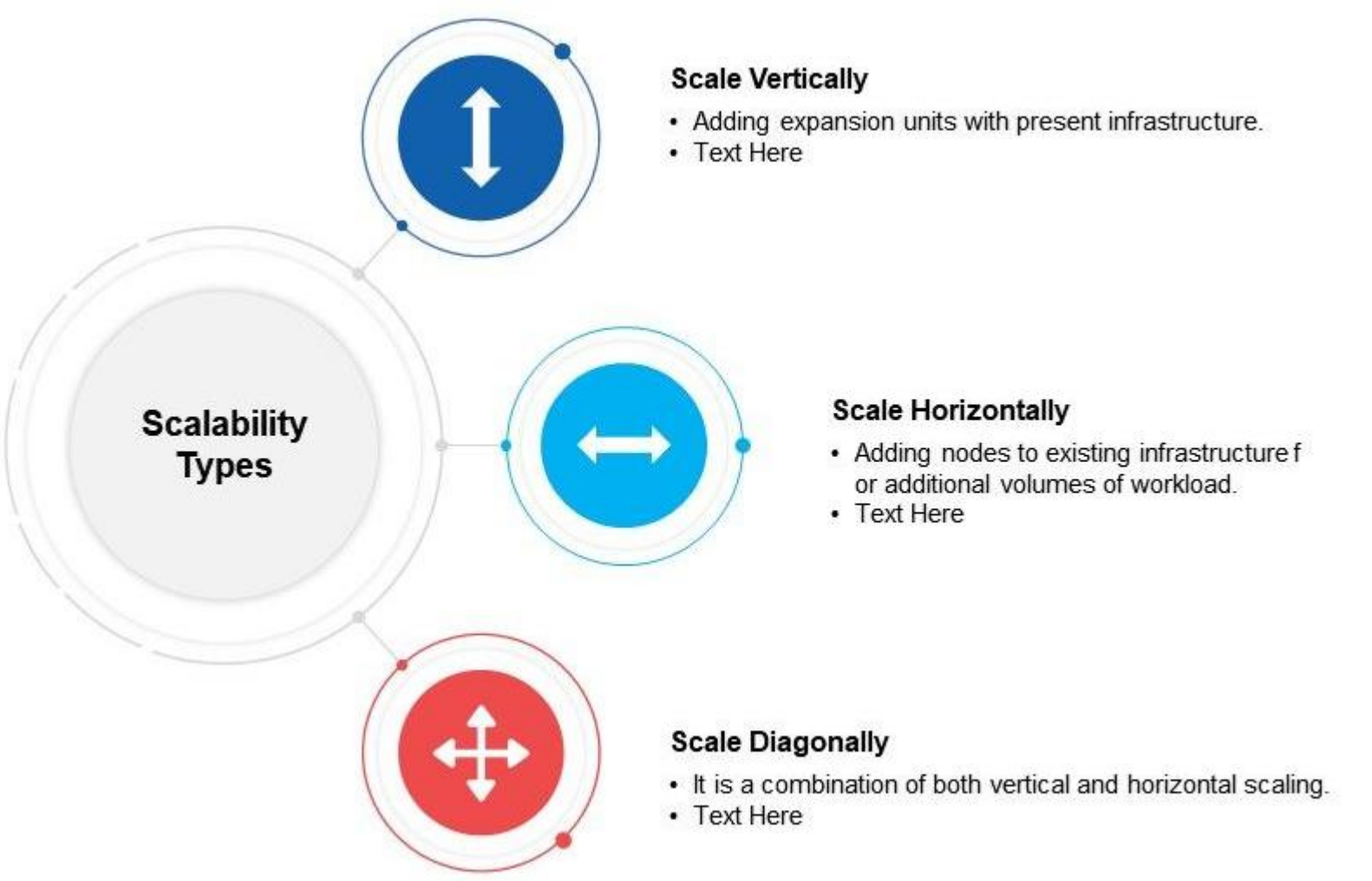
Vertical Scaling:

- Scaling up a single node
  - Physical limitations – instances are very powerful but still have finite limits
  - Resources such as number of sockets can only go so high

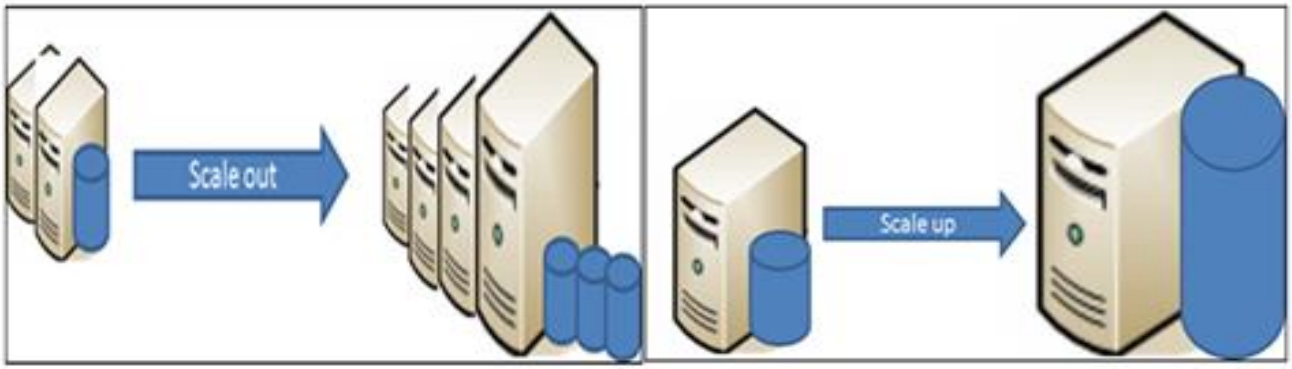
Horizontal Scaling:

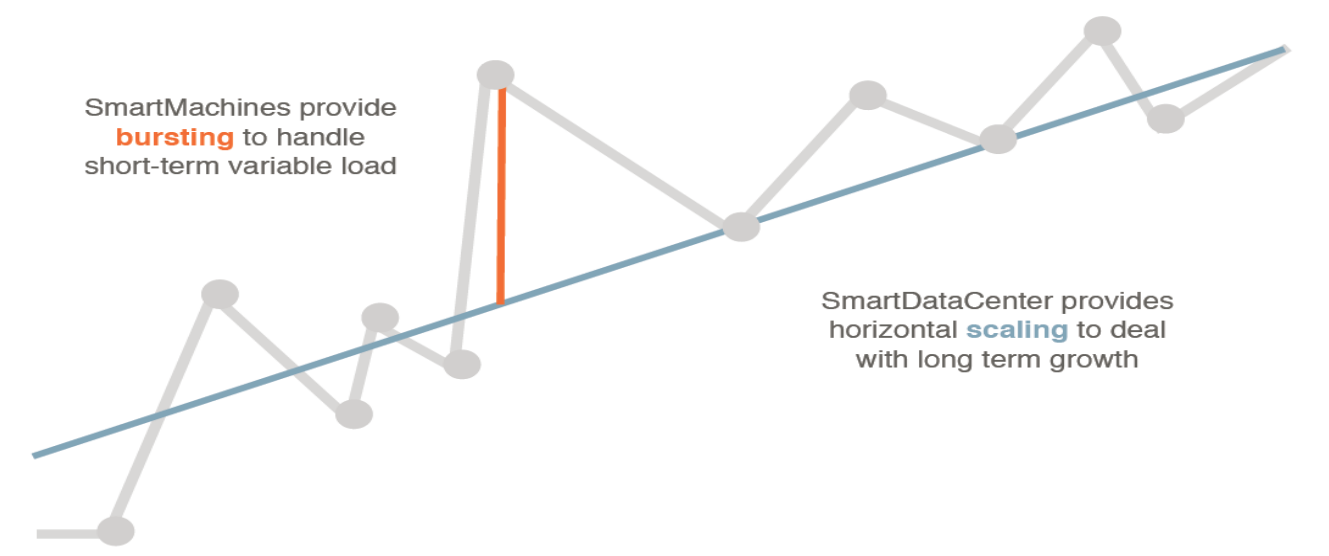
- Scaling out across multiple nodes
  - Ability to distribute traffic over a number of nodes
  - Allows for more flexibility over time

Three Major Types Scalability in Cloud Computing



Vertical vs. Horizontal



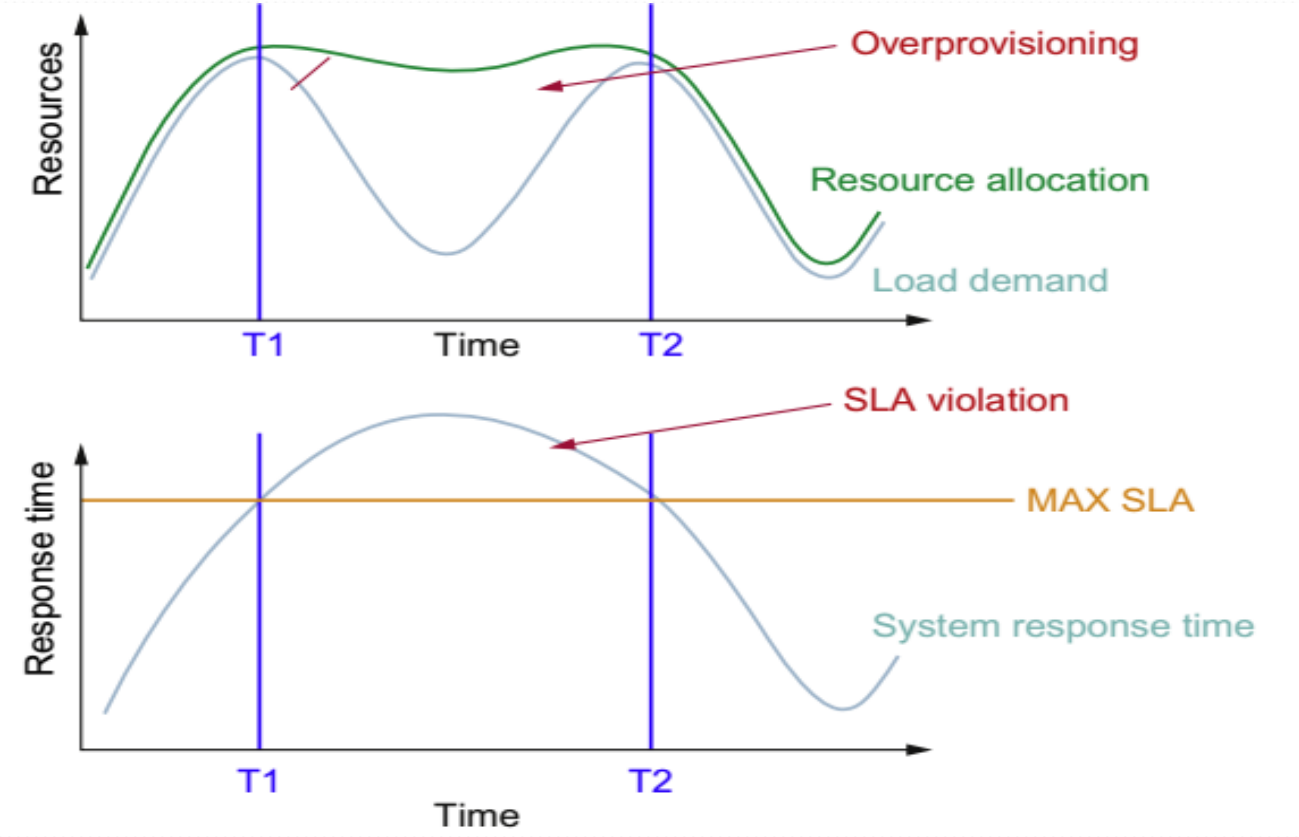


Auto-Scaling

- By dynamically adjusting the amount of resources selected for service applications without human intervention, it is possible to provide computational and storage demands of workload fluctuations, which prevent termination of application services.
- Dynamic resource provisioning is a solution for improving the utilization of cloud resources against application services and user objectives.
- Auto-scaling (ability to scale-up and scale-down VMs resources) avoid over-provisioning (leads to cloud resource wastage and increase the cost of resources) and under-provisioning (causes performance degradation and violation of service level agreement) with a prediction of resources automatically.

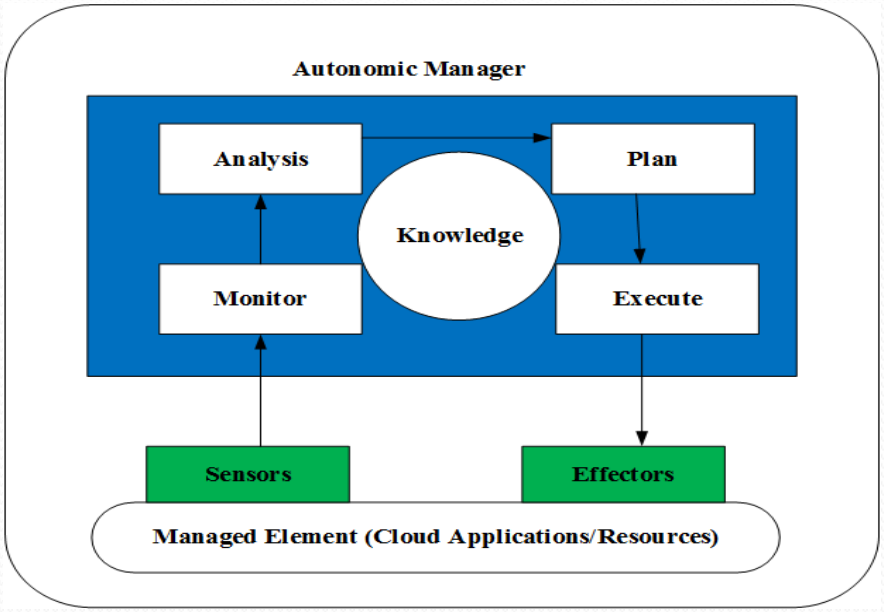
Auto-Scaler Challenges:

- **Over-Provisioning**
- **Under-Utilization**
- **Ociliation**



Autonomic Computing

- To effectively manage cloud deployed applications it is essential that reactions to regularly occurring or anticipated events are built into the system.
- The autonomic computing can be applied to implement a cloud application system which knows its state and reacts to changes in it.
- The term autonomic computing was first used by IBM in 2001 to describe computing systems that are said to be self-managing




Clouds
Cloud Services
Cloud Consumers