IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES

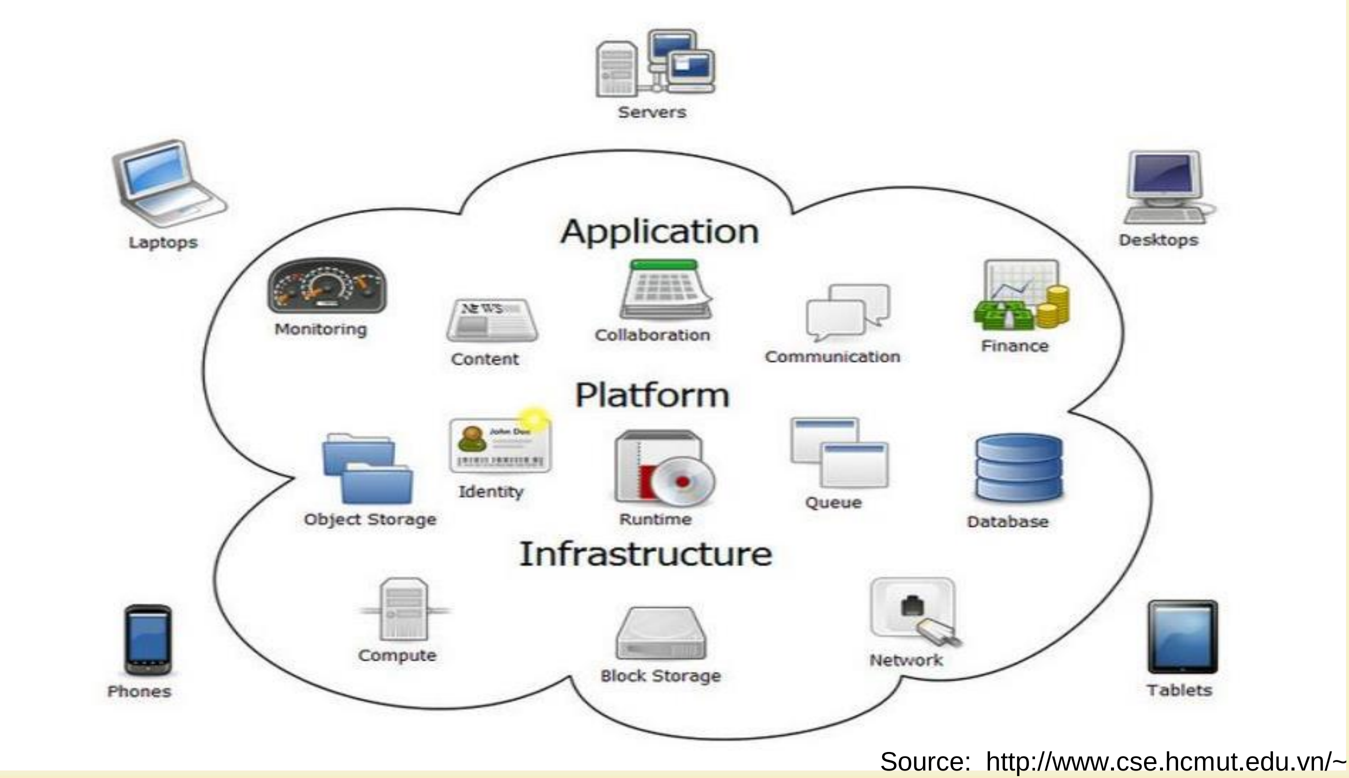
NPTEL

# CLOUD COMPUTING

## Resource Management - II

PROF. SOUMYA K. GHOSH
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

# Different Resources in Computing



Source: http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf

# Resources types

- **Physical resource**
  - ❑ Computer, disk, database, network, scientific instruments.
- **Logical resource**
  - ❑ Execution, monitoring, communicate application .

Source:  http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf

# Resources Management

- The term ***resource management*** refers to the operations used to control how capabilities provided by Cloud resources and services cane be made available to other entities, whether users, applications, services in an *efficient* manner.

Source: http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf

# Resource Management for IaaS

- Infrastructure-as-a-Service (IaaS) is most popular cloud service
- In IaaS, cloud providers offer resources that include computers as virtual machines, raw (block) storage, firewalls, load balancers, and network devices.
- One of the major challenges in IaaS is resource management.

Source: http://www.zearon.com/down/Resource%20management%20for%20Infrastructure%20as%20a%20Service%20%28IaaS%29%20in%20cloud%20computing%20A%20survey.pdf

# Resource Management - Objectives

- Scalability
- Quality of service
- Optimal utility
- Reduced overheads
- Improved throughput
- Reduced latency
- Specialized environment
- Cost effectiveness
- Simplified interface

# Resource Management – Challenges (Hardware)

- CPU (central processing unit)
- Memory
- Storage
- Workstations
- Network elements
- Sensors/actuators

# **Resource Management – Challenges** (Logical resources)

- Operating system
- Energy
- Network throughput/bandwidth
- Load balancing mechanisms
- Information security
- Delays
- APIs/(Applications Programming Interfaces)
- Protocols

# Resource Management Aspects

- Resource provisioning
- Resource allocation
- Resource requirement mapping
- Resource adaptation
- Resource discovery
- Resource brokering
- Resource estimation
- Resource modeling

# Resource Management

| Type | Details |
|---|---|
| **Resource provisioning** | Allocation of a service provider's resources to a customer |
| **Resource allocation** | Distribution of resources economically among competing groups of people or programs |
| **Resource adaptation** | Ability or capacity of that system to adjust the resources dynamically to fulfill the requirements of the user |
| **Resource mapping** | Correspondence between resources required by the users and resources available with the provider |
| **Resource modeling** | Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network.<br>Attributes of resource management: states, transitions, inputs and outputs within a given environment.<br>Resource modeling helps to predict the resource requirements in subsequent time intervals |
| **Resource estimation** | A close guess of the actual resources required for an application, usually with some thought or calculation involved |
| **Resource discovery and selection** | Identification of list of authenticated resources that are available for job submission and to choose the best among them |
| **Resource brokering** | It is the negotiation of the resources through an agent to ensure that the necessary resources are available at the right time to complete the objectives |
| **Resource scheduling** | A resource schedule is a timetable of events and resources. Shared resources are available at certain times and events are planned during these times. In other words, It is determining when an activity should start or end, depending on its (1) duration, (2) predecessor activities, (3) predecessor relationships, and (4) resources allocated |

# Resource Provisioning Approaches

| Approach | Description |
|---|---|
| **Nash equilibrium approach using Game theory** | Run time management and allocation of IaaS resources considering several criteria such as the heterogeneous distribution of resources, rational exchange behaviors of cloud users, incomplete common information and dynamic successive allocation |
| **Network queuing model** | Presents a model based on a network of queues, where the queues represent different tiers of the application.The model sufficiently captures the behavior of tiers with significantly different performance characteristics and application idiosyncrasies, such as, session-based workloads, concurrency limits, and caching at intermediate tiers |
| **Prototype provisioning** | Employs the k-means clustering algorithm to automatically determine the workload mix and a queuing model to predict the server capacity for a given workload mix. |
| **Resource (VM) provisioning** | Uses virtual machines (VMs) that run on top of the Xen hypervisor. The system provides a Simple Earliest Deadline First (SEDF) scheduler that implements weighted fair sharing of the CPU capacity among all the VMs<br>The share of CPU cycles for a particular VM can be changed at runtime |
| **Adaptive resource provisioning** | Automatic bottleneck detection and resolution under dynamic resource management which has the potential to enable cloud infrastructure providers to provide SLAs for web applications that guarantee specific response time requirements while minimizing resource utilization. |
| **SLA oriented methods** | Handling the process of dynamic provisioning to meet user SLAs in autonomic manner. Additional resources are provisioned for applications when required and are removed when they are not necessary |
| **Dynamic and automated framework** | A dynamic and automated framework which can adapt the adaptive parameters to meet the specific accuracy goal, and then dynamically converge to near-optimal resource allocation to handle unexpected changes |
| **Optimal cloud resource provisioning (OCRP)** | The demand and price uncertainty is considered using optimal cloud resource provisioning (OCRP) including deterministic equivalent formulation, sample-average approximation, etc. |

# Resource Allocation Approaches

| Approach | Description |
|---|---|
| **Market-oriented resource allocation** | Considers the case of a single cloud provider and address the question how to best match customer demand in terms of both supply and price in order to maximize the providers revenue and customer satisfactions while minimizing energy cost. In particular, it models the problem as a constrained discrete-time optimal control problem and uses Model Predictive Control(MPC) to find its solution |
| **Intelligent multi-agent model** | An intelligent multi-agent model based on virtualization rules for resource virtualization to automatically allocate service resources suitable for mobile devices. It infers user demand by analyzing and learning user context information. |
| **Energy-Aware Resource allocation** | Resource allocation is carried out by mimicking the behavior of ants, that the ants are likely to choose the path identified as a shortest path, which is indicated by a relatively higher density of pheromone left on the path compared to other possible paths |
| **Measurement based analysis on performance** | Focuses on measurement based analysis on performance impact of co-locating applications in a virtualized cloud in terms of throughput and resource sharing effectiveness, including the impact of idle instances on applications that are running concurrently on the same physical host |
| **Dynamic resource allocation method** | Dynamic resource allocation method based on the load of VMs on IaaS, which enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user |
| **Real time resource allocation mechanism** | Designed for helping small and medium sized IaaS cloud providers to better utilize their hardware resources with minimum operational cost by a well-designed underlying hardware infrastructure, an efficient resource scheduling algorithm and a set of migrating operations of VMs |
| **Dynamic scheduling and consolidation mechanism** | Presents the architecture and algorithmic blueprints of a framework for workload co-location, which provides customers with the ability to formally express workload scheduling flexibilities using Directed Acyclic Graphs (DAGs), and optimizes the use of cloud resources to collocate client's workloads |

# Resource Mapping Approaches

| Approach | Description |
|---|---|
| **Symmetric mapping pattern** | Symmetric mapping pattern for the design of resource supply systems. It divides resource supply in three functions: (1) users and providers match and engage in resource supply agreements, (2) users place tasks on subscribed resource containers, and (3) providers place supplied resource containers on physical resources |
| **Load-aware mapping** | Explores how to simplify VM image management and reduce image preparation overhead by the multicast file transferring and image caching/reusing. Load-Aware Mapping to further reduce deploying overhead and make efficient use of resources. |
| **Minimum congestion mapping** | Framework for solving a natural graph mapping problem arising in cloud computing. Applying this framework to obtain offline and online approximation algorithms for workloads given by depth-d trees and complete graphs |
| **Iterated local search based request partitioning** | Request partitioning approach based on iterated local search is introduced that facilitates the cost- efficient and on-line splitting of user requests among eligible Cloud Service Providers (CSPs) within a networked cloud environment |
| **SOA API** | Designed to accept different resource usage prediction models and map QoS constraints to resources from various IaaS providers |
| **Impatient task mapping** | Batch mapping via genetic algorithms with throughput as a fitness function that can be used to map jobs to cloud resources |
| **Distributed ensembles of virtual appliances (DEVAs)** | Requirements are inferred by observing the behavior of the system under different conditions and creating a model that can be later used to obtain approximate parameters to provide the resources. |
| **Mapping a virtual network onto a substrate network** | An effective method (using backbone mapping) for computing high quality mappings of virtual networks onto substrate networks. The computed virtual networks are constructed to have sufficient capacity to accommodate any traffic pattern allowed by user-specified traffic constraints. |

# Resource Adaptation Approaches

| Approach | Description |
|---|---|
| **Reinforcement learning guided control policy** | A multi-input multi-output feedback control model-based dynamic resource provisioning algorithm which adopts reinforcement learning to adjust adaptive parameters to guarantee the optimal application benefit within the time constraint |
| **Web-service based prototype** | A web-service based prototype framework, and used it for performance evaluation of various resource adaptation algorithms under different realistic settings |
| **OnTimeMeasure service** | Presents an application – adaptation case study that uses OnTimeMeasure-enabled performance intelligence in the context of dynamic resource allocation within thin-client based virtual desktop clouds to increase cloud scalability, while simultaneously delivering satisfactory user quality-of-experience |
| **Virtual networks** | Proposes virtual networks architecture as a mechanism in cloud computing that can aggregate traffic isolation, improving security and facilitating pricing, also allowing customers to act in cases where the performance is not in accordance with the contract for services |
| **DNS-based Load Balancing** | Proposes a system that contain the appropriate elements so that applications can be scaled by replicating VMs (or application containers), by reconfiguring them on the fly, and by adding load balancers in front of these replicas that can scale by themselves |
| **Hybrid approach** | Proposes a mechanism for providing dynamic management in virtualized consolidated server environments that host multiple multi-tier applications using layered queuing models for Xen-based virtual machine environments, which is a novel optimization technique that uses a combination of bin packing and gradient search |

# Performance Metrics for Resource Management

- Reliability
- Ease of deployment
- QoS
- Delay
- Control overhead

# Thank you!