CS6006 - CLOUD COMPUTING

Module 8 - Cloud Management

Presented By
Dr. S. Muthurajkumar,
Assistant Professor,
Dept. of CT, MIT Campus,
Anna University, Chennai

CLOUD MANAGEMENT

- Administrating the Clouds
- Management Responsibilities & lifecycle Management
- Distributed Management of Virtual Infrastructures
- SLA – An Inspiration
- Traditional Approaches to SLO Management
- Types of SLA
- Life Cycle of SLA
- SLA Management in Cloud
- Automated Policy-based Management
- Cloud Management Standards

ADMINISTRATING THE CLOUDS

- The explosive growth in cloud computing services has led many vendors to rename their products and reposition them to get in on the gold rush in the clouds.
- What was once a network management product is now a cloud management product.
- Nevertheless, this is one area of technology that is very actively funded, comes replete with interesting startups, has been the focus of several recent strategic acquisitions, and has resulted in some interesting product alliances.

ADMINISTRATING THE CLOUDS

- **These fundamental features are offered by traditional network management systems:**
- Administration of resources
- Configuring resources
- Enforcing security
- Monitoring operations
- Optimizing performance
- Policy management
- Performing maintenance
- Provisioning of resources

ADMINISTRATING THE CLOUDS

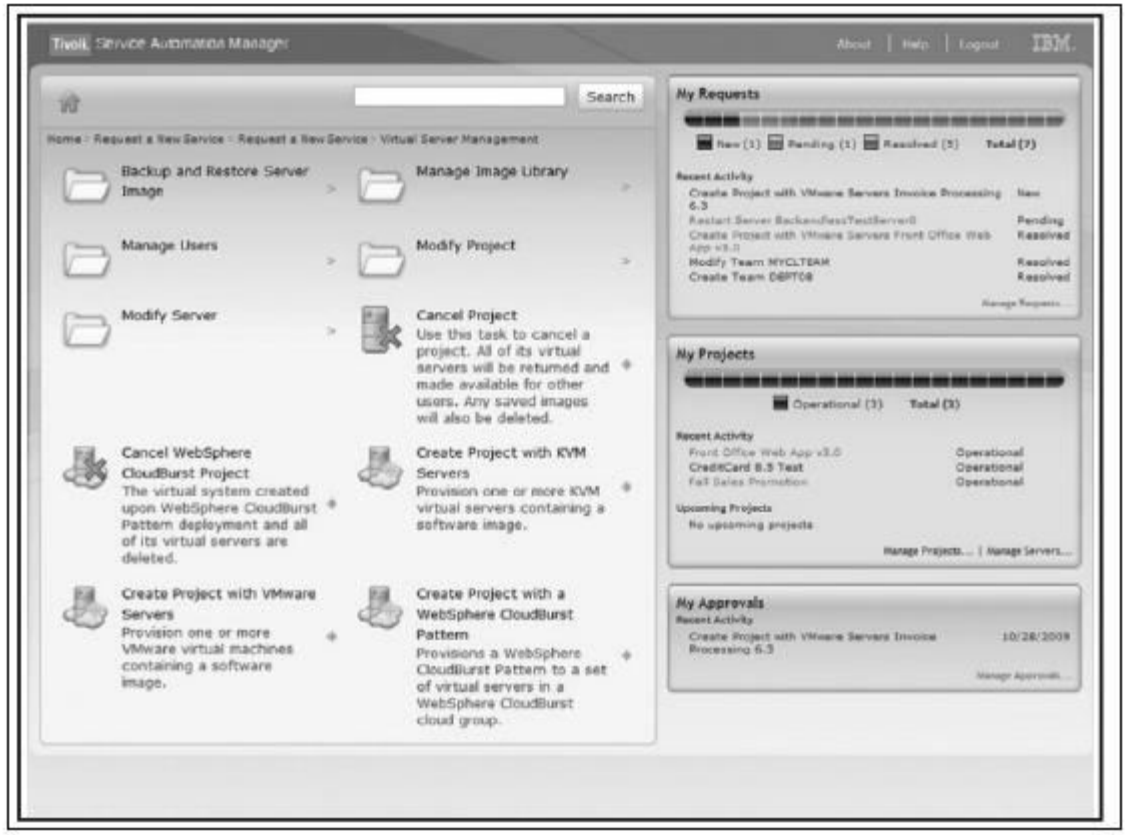
- **Network management systems are often described in terms of the acronym FCAPS, which stands for these features:**
- Fault
- Configuration
- Accounting
- Performance
- Security
- Most network management packages have one or more of these characteristics; no single package provides all five elements of FCAPS.

ADMINISTRATING THE CLOUDS

- The cloud products for these five large cloud vendors at the following URLs:
- BMC Cloud Computing (http://www.bmc.com/solutions/esm-initiative/cloud-computing.html)
- Computer Associates Cloud Solutions (http://www.ca.com/us/cloudcomputing.aspx)
- HP Cloud Computing (http://h20338.www2.hp.com/enterprise/w1/en/technologies/cloud-computing-overview.html)
- IBM Cloud Computing (http://www.ibm.com/ibm/cloud/)
- Microsoft Cloud Services (http://www.microsoft.com/cloud/)

ADMINISTRATING THE CLOUDS

- IBM Tivoli Service Automation Manager, a framework tool for managing cloud infrastructure



MANAGEMENT RESPONSIBILITIES

- What separates a network management package from a cloud computing management package is the “cloudly” characteristics that cloud management service must have:
- Billing is on a pay-as-you-go basis.
- The management service is extremely scalable.
- The management service is ubiquitous.
- Communication between the cloud and other systems uses cloud networking standards.

MANAGEMENT RESPONSIBILITIES

- **To monitor an entire cloud computing deployment stack, you monitor six different categories:**

1. End-user services such as HTTP, TCP, POP3/SMTP, and others
2. Browser performance on the client
3. Application monitoring in the cloud, such as Apache, MySQL, and so on
4. Cloud infrastructure monitoring of services such as Amazon Web Services, GoGrid, Rackspace, and others
5. Machine instance monitoring where the service measures processor utilization, memory usage, disk consumption, queue lengths, and other important parameters

MANAGEMENT RESPONSIBILITIES

- **To monitor an entire cloud computing deployment stack, you monitor six different categories:**

6. Network monitoring and discovery using standard protocols like the Simple Network Management Protocol (SNMP), Configuration Management Database (CMDB), Windows Management Instrumentation (WMI), and the like

It's important to note that there are really two aspects to cloud management:

1. Managing resources in the cloud

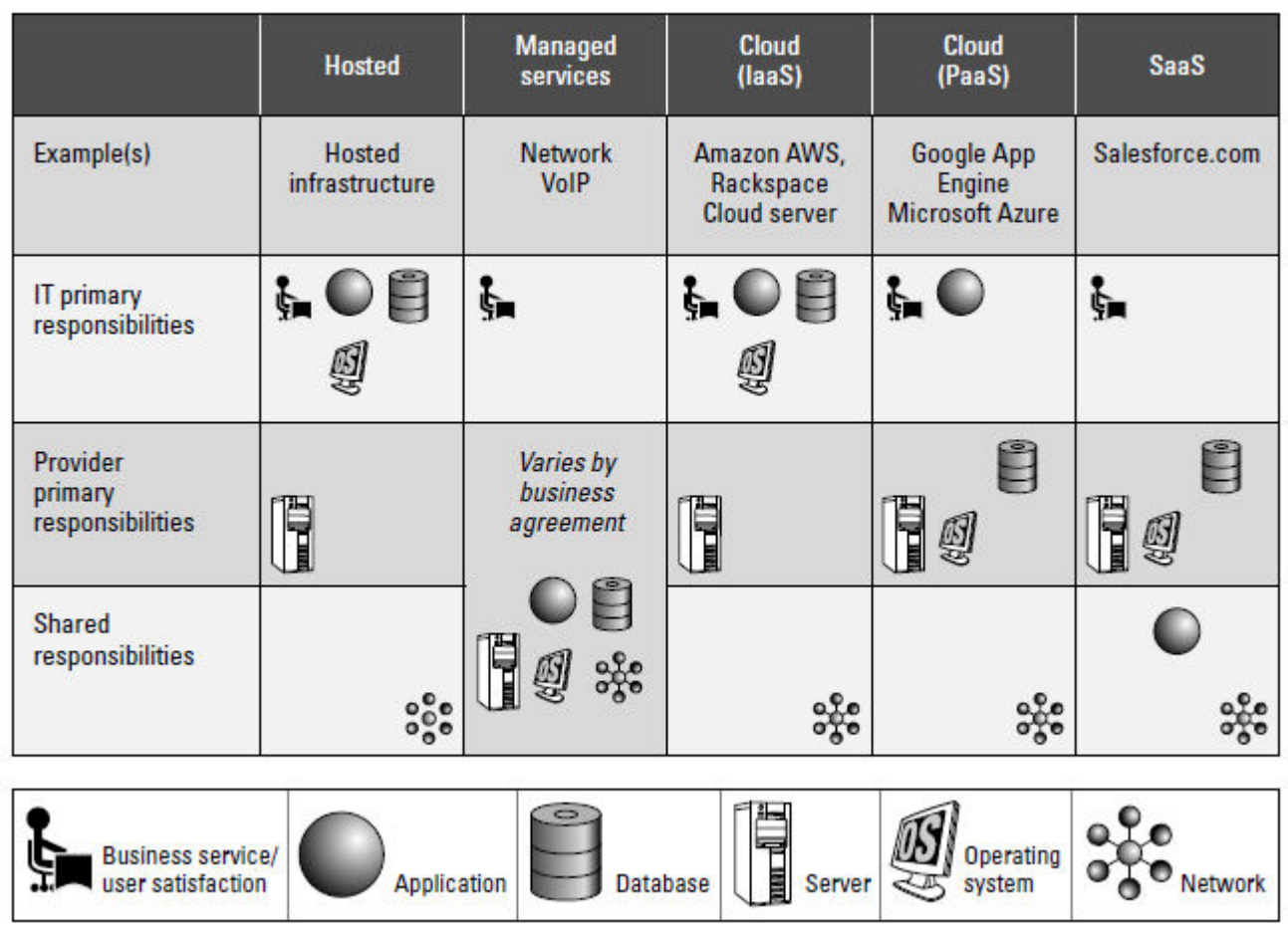
2. Using the cloud to manage resources on-premises

MANAGEMENT RESPONSIBILITIES

- Platform as a Service (PaaS) – Windows Azure or Google App Engine
- Software as a Service (SaaS) – Salesforce.com
- **The following monitoring capabilities:**
- Create a new application, and set it up in your domain.
- Invite other people to be part of developing your application.
- View data and error logs.
- Analyze your network traffic.
- Browse the application datastore, and manage its indexes.
- View the application's scheduled tasks.
- Test the application, and swap out versions.

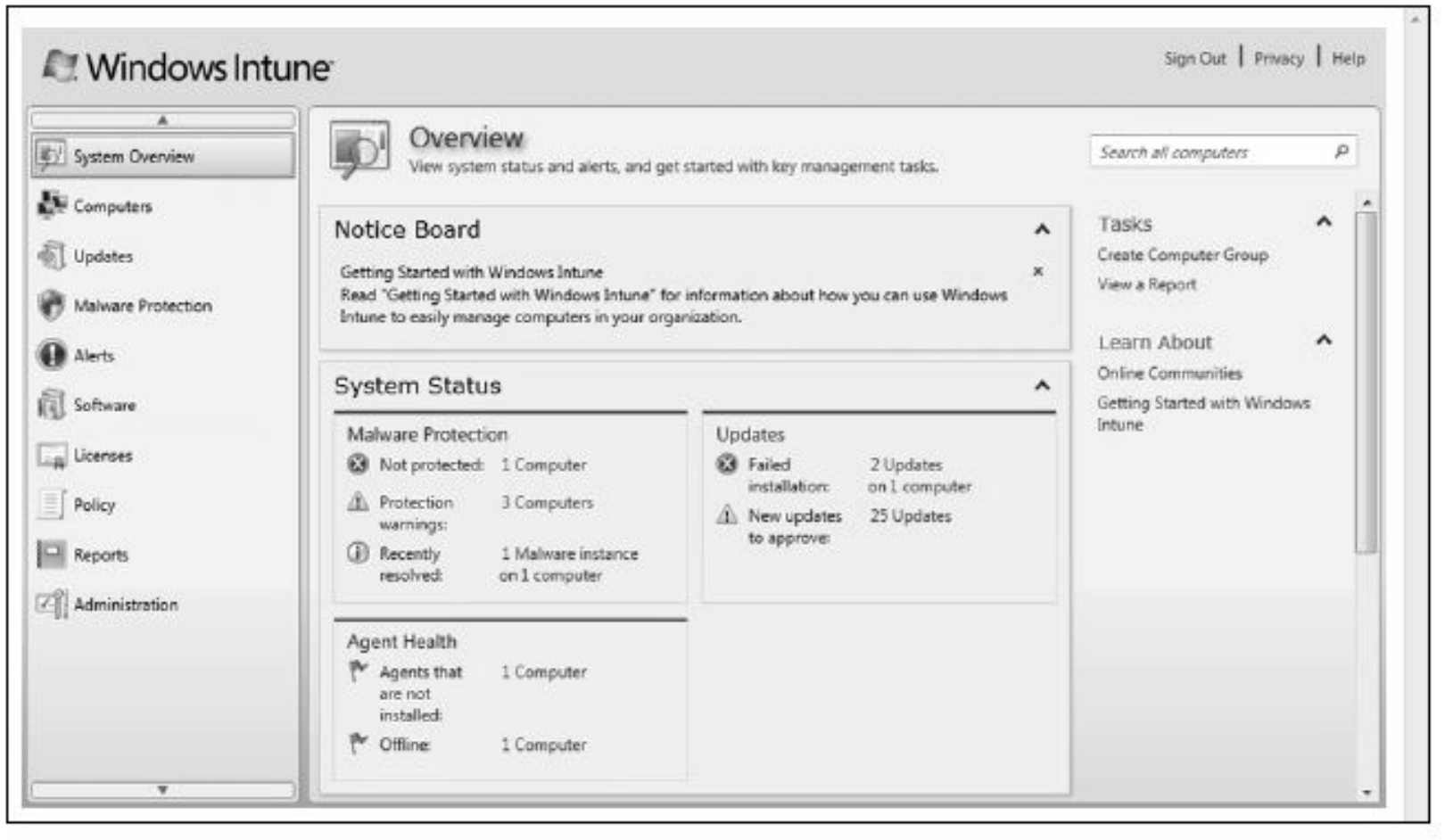
MANAGEMENT RESPONSIBILITIES

- Management responsibilities by service model type



MANAGEMENT RESPONSIBILITIES

- Intune is Microsoft's cloud-based management service for Windows systems



LIFECYCLE MANAGEMENT

- Cloud services have a defined lifecycle, just like any other system deployment. A management program has to touch on each of the six different stages in that lifecycle:

1. The definition of the service as a template for creating instances Tasks performed in Phase 1 include the creation, updating, and deletion of service templates.

2. Client interactions with the service, usually through an SLA (Service Level Agreement) contract

This phase manages client relationships and creates and manages service contracts.

3. The deployment of an instance to the cloud and the runtime management of instances Tasks performed in Phase 3 include the creation, updating, and deletion of service offerings.

4. The definition of the attributes of the service while in operation and performance of modifications of its properties The chief task during this management phase is to perform service optimization and customization.

LIFECYCLE MANAGEMENT

- Cloud services have a defined lifecycle, just like any other system deployment. A management program has to touch on each of the six different stages in that lifecycle:

5. Management of the operation of instances and routine maintenance

During Phase 5, you must monitor resources, track and respond to events, and perform reporting and billing functions.

6. Retirement of the service End of life tasks include data protection and system migration, archiving, and service contract termination.

CLOUD MANAGEMENT PRODUCTS

- Cloud management software and services is a very young industry, and as such, it has a very large number of companies, some with new products and others with older products competing in this area.

Product	URL	Description
AbiCloud	http://www.abiquo.com/	Virtual machine conversion and management
Amazon CloudWatch	http://aws.amazon.com/cloudwatch/	AWS dashboard
BMC Cloud Computing Initiative	http://www.bmc.com/solutions/esm-initiative/cloud-computing.html	Cloud planning, lifecycle management, optimization, and guidance
CA Cloud Connected Management Suite	http://www.ca.com/us/cloud-solutions.aspx	CA Cloud Insight, CA Cloud Compose, CA Cloud Optimize, and CA Cloud Orchestrate are described below
Cacti	http://www.cacti.net/	Network performance graphing solution
CloudKick	https://www.cloudkick.com/	Cloud server monitoring
Dell Scalent	http://www.scalent.com/index.php	Virtualization provisioning system that will be rolled into Dell's Advanced Infrastructure Manager (AIM)
Elastra	http://www.elastra.com/	Federated hybrid cloud management software
Ganglia	http://ganglia.info/	Distributed network monitoring software
Gomez	http://www.gomez.com/	Web site monitoring and analytics
HP Cloud Computing	http://h20338.www2.hp.com/enterprise/wl/en/technologies/cloud-computing-overview.html	A variety of management products and services, both released and under development
Hyperic	http://www.hyperic.com/	Performance management for virtualized Java Apps with VMware integration
IBM Service Management and Cloud Computing	http://www-01.ibm.com/software/tivoli/solutions/cloudcomputing/	Various IBM Tivoli managers and monitors
Internetseer	http://www.internetseer.com/home/index.xtp	Web site monitoring service
Intune	http://www.microsoft.com/windows/windowsintune/default.aspx	Cloud-based Windows system management
Keynote	http://www.keynote.com/	Web, mobile, streaming, and customer test and measurement products

DISTRIBUTED MANAGEMENT OF VIRTUAL INFRASTRUCTURES

- The problem of efficiently selecting or scheduling computational resources is well known. However, the state of the art in VM-based resource scheduling follows a static approach, where resources are initially selected using a greedy allocation strategy, with minimal or no support for other placement policies.
- To efficiently schedule resources, VI managers must be able to support flexible and complex scheduling policies and must leverage (use) the ability of VMs to suspend, resume, and migrate.
- This complex task is one of the core problems that the RESERVOIR project tries to solve. The problem of how to manage VMs distributed across a pool of physical resources will be described. OpenNebula, the virtual infrastructure manager developed by the RESERVOIR project will be explained later.

DISTRIBUTED MANAGEMENT OF VIRTUAL INFRASTRUCTURES

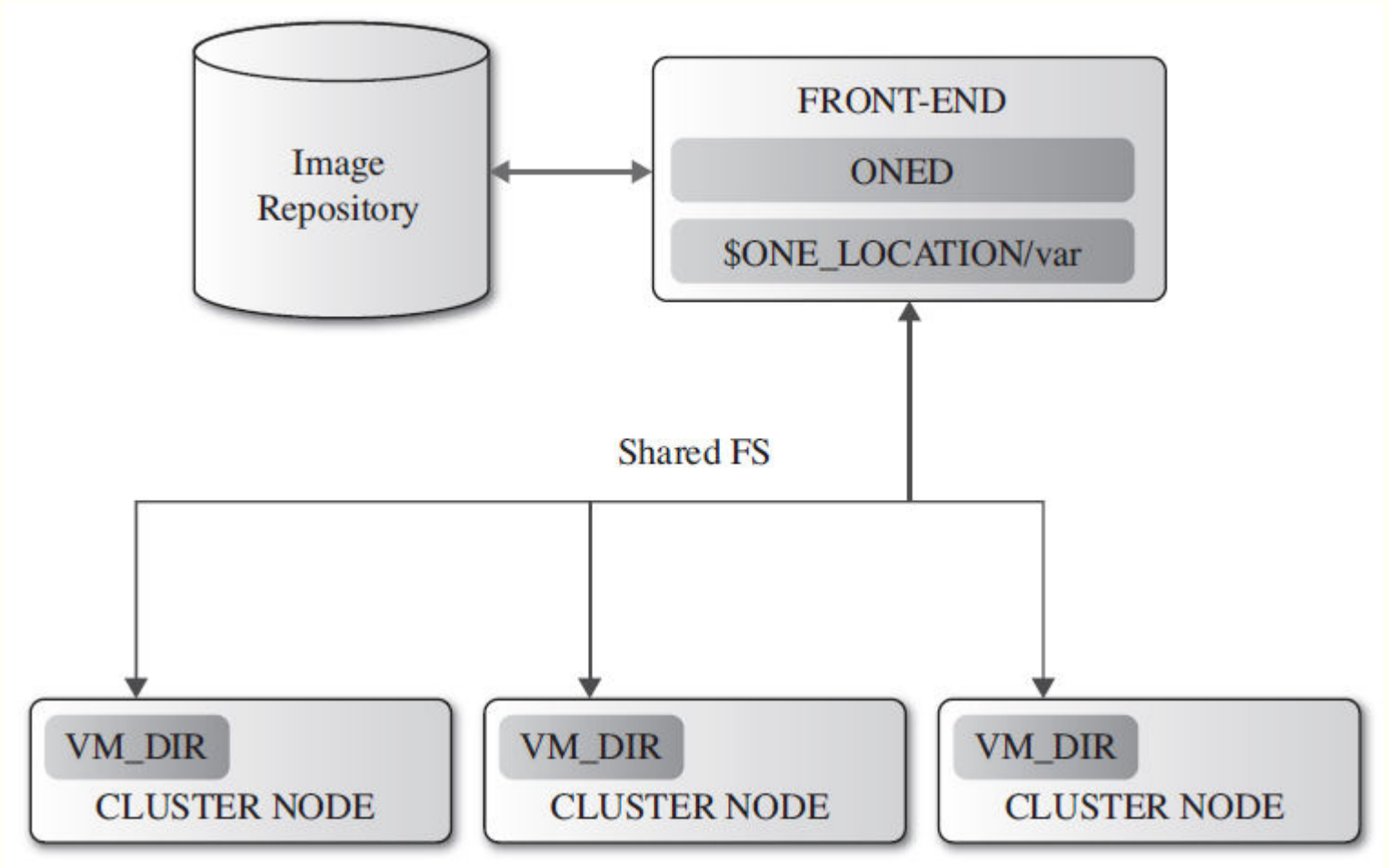
- Virtual Infrastructure (VI) management—the management of virtual machines distributed across a pool of physical resources—becomes a key concern when building an IaaS cloud and poses a number of challenges. configuration, including preparation of the machine's software environment and network configuration.
- The problem of efficiently selecting or scheduling computational resources is well known. However, the state of the art in VM-based resource scheduling follows a static approach, where resources are initially selected using a greedy allocation strategy, with minimal or no support for other placement policies.
- To efficiently schedule resources, VI managers must be able to support flexible and complex scheduling policies and must leverage (use) the ability of VMs to suspend, resume, and migrate.
- This complex task is one of the core problems that the RESERVOIR project tries to solve. The problem of how to manage VMs distributed across a pool of physical resources will be described. OpenNebula, the virtual infrastructure manager developed by the RESERVOIR project will be explained later.

DISTRIBUTED MANAGEMENT OF VIRTUAL INFRASTRUCTURES

- Managing VMs in a pool of distributed physical resources is a key concern in IaaS clouds, requiring the use of a virtual infrastructure manager.
- Example:
- OpenNebula is capable of managing groups of interconnected VMs—with support for the Xen, KVM, and VMWare platforms—within data centers and private clouds that involve a large amount of virtual and physical servers.
- OpenNebula can also be used to build hybrid clouds by interfacing with remote cloud sites.
- VM Model and Life Cycle
- VM Management
- Further Reading on OpenNebula

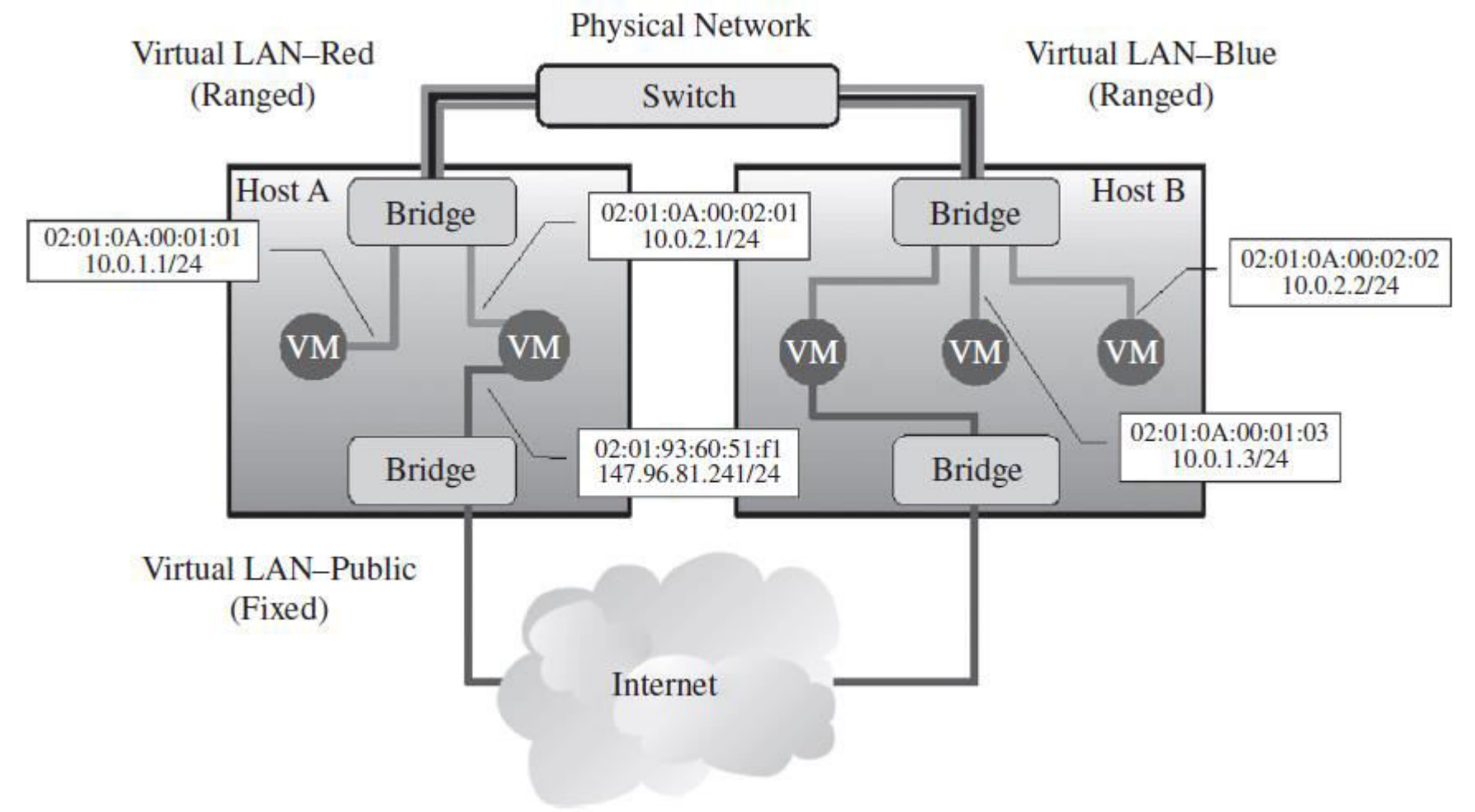
DISTRIBUTED MANAGEMENT OF VIRTUAL INFRASTRUCTURES

- Image management in OpenNebula



DISTRIBUTED MANAGEMENT OF VIRTUAL INFRASTRUCTURES

- Networkig model for OpenNebula

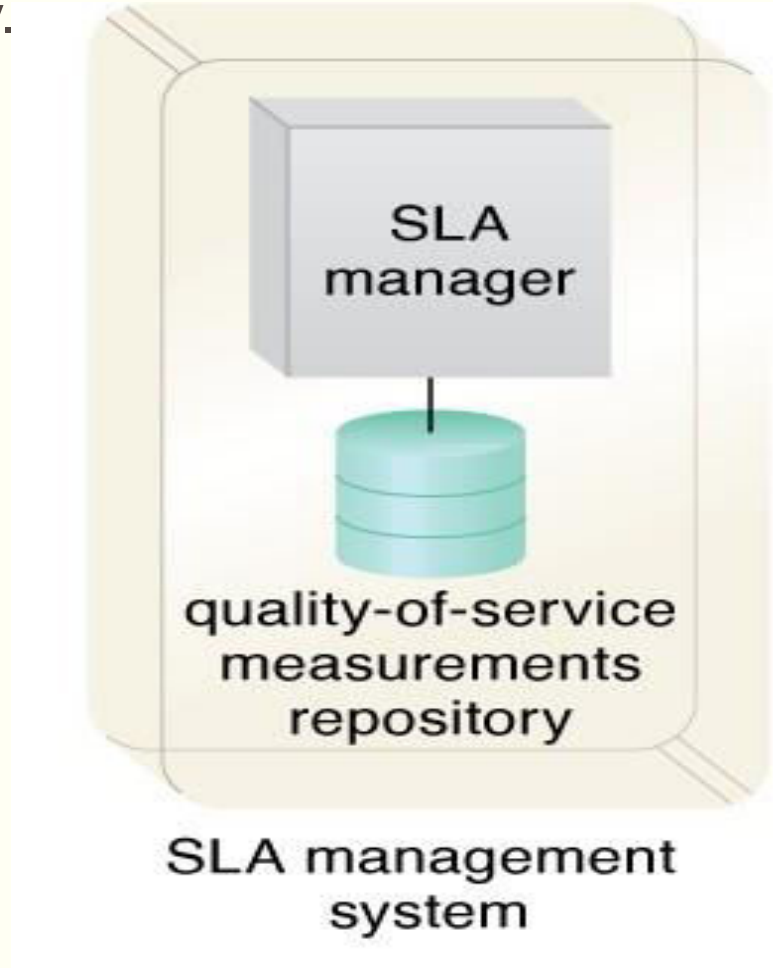


SLA MANAGEMENT SYSTEM

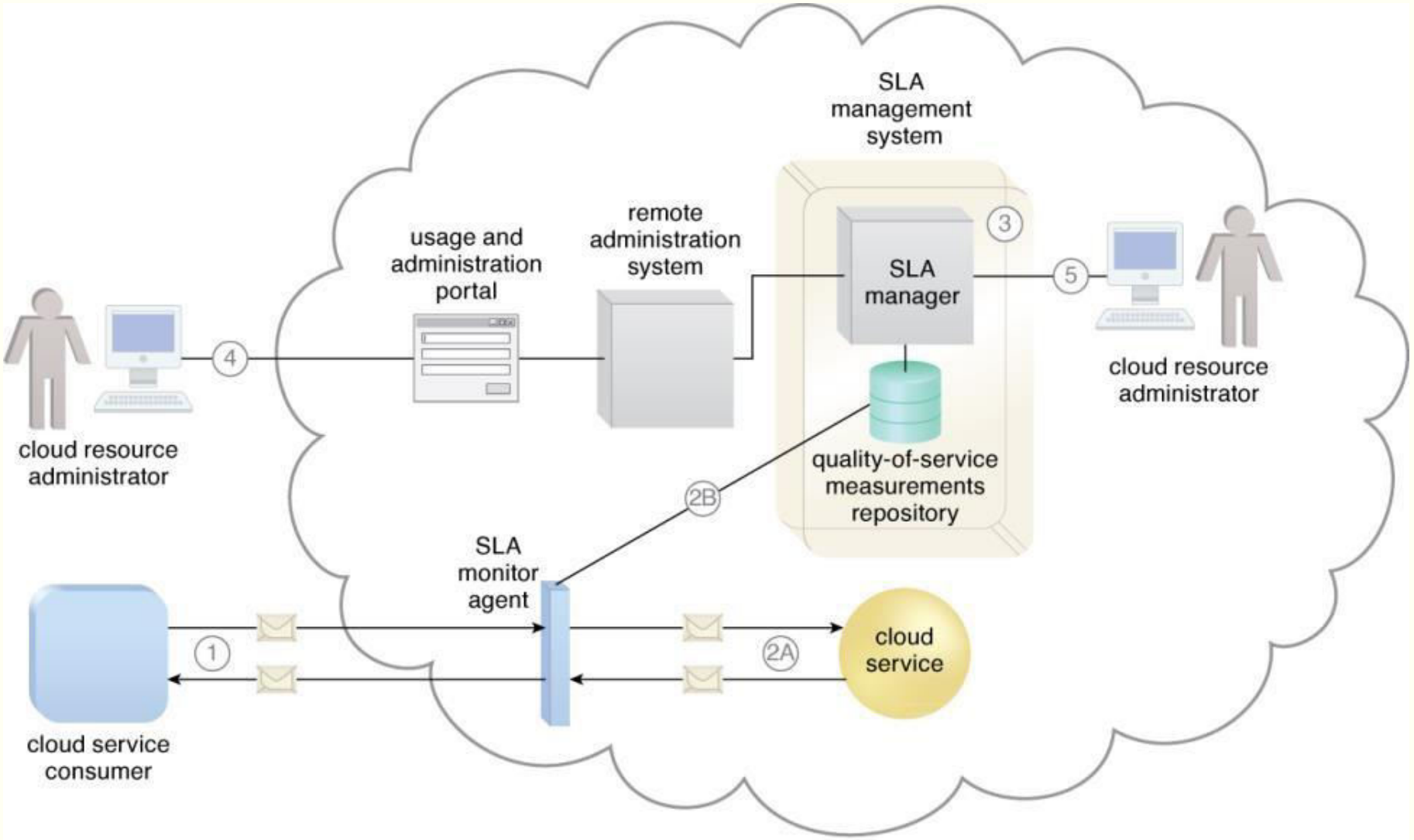
- The SLA management system mechanism represents a range of commercially available cloud management products that provide features pertaining to the administration, collection, storage, reporting, and runtime notification of SLA data.
- An SLA management system deployment generally include a repository used to store and retrieve collected SLA data based on pre-defined metrics and reporting parameters.

SLA MANAGEMENT SYSTEM

- An SLA management system encompassing an SLA manager and QoS measurements repository.



SLA MANAGEMENT SYSTEM



SLA MANAGEMENT SYSTEM

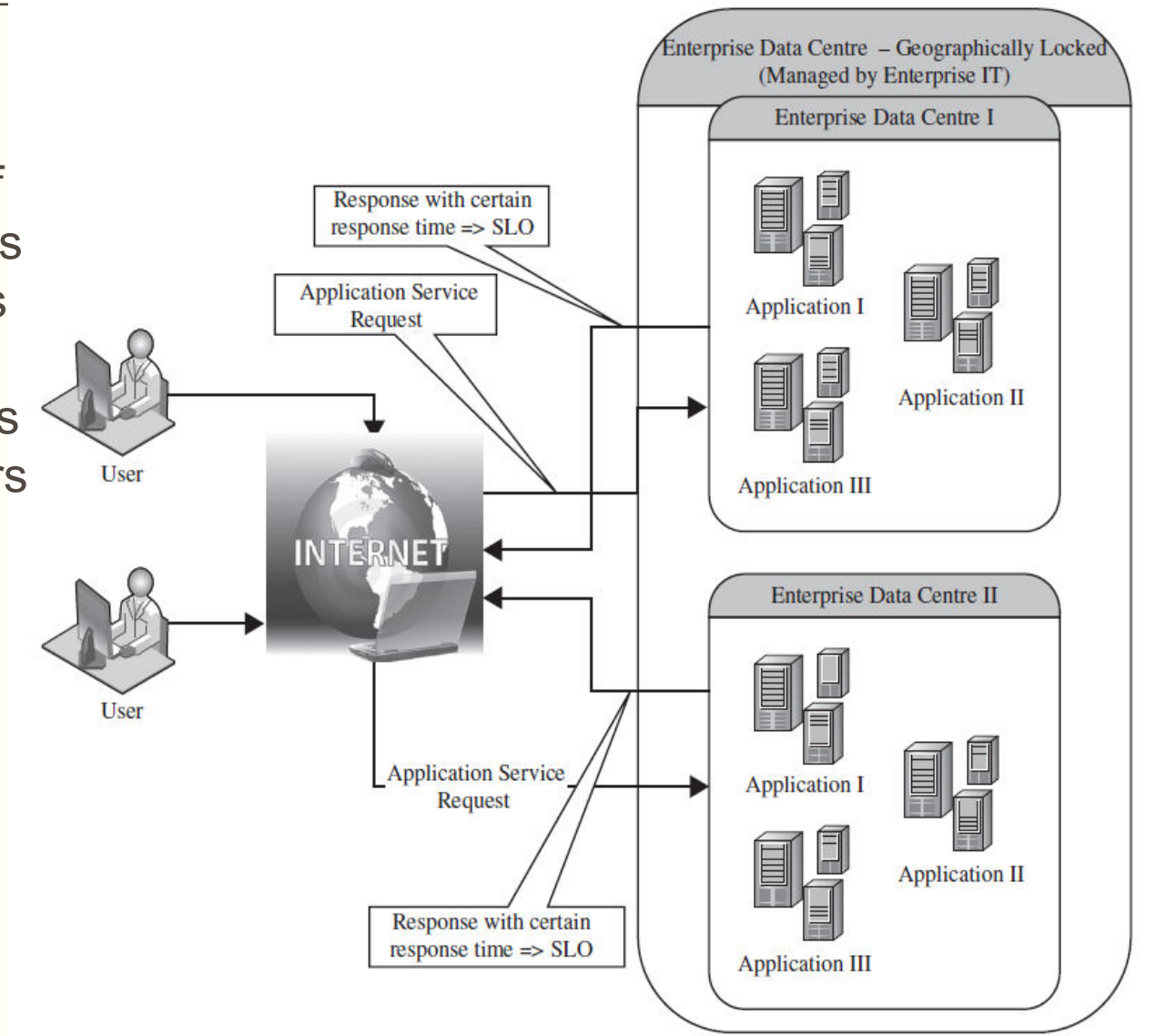
- A cloud service consumer interacts with a cloud service (1).
- An SLA monitor intercepts the exchanged messages, evaluates the interaction, and collects relevant runtime data in relation to quality-of-service guarantees defined in the cloud service's SLA (2A).
- The data collected is stored in a repository (2B) that is part of the SLA management system (3).
- Queries can be issued and reports can be generated for a cloud resource administrator with an external cloud consumer via a usage and administrator portal (4) or for an internal cloud resource administrator via the SLA management system's native user interface (5).

AN INSPIRATION

- Provisioning in those days involved deciding hardware configuration, determining the number of physical machines, and acquiring them upfront so that the overall business objectives could be achieved.
- The web applications were hosted on these dedicated individual servers within enterprises' own server rooms.
- These web applications were used to provide different kinds of e-services to various clients.
- Typically, the service-level objectives (SLOs) for these applications were response time and throughput of the application end-user requests.
- The capacity buildup was to cater to the estimated peak load experienced by the application.
- The activity of determining the number of servers and their capacity that could satisfactorily serve the application end-user requests at peak loads is called **capacity planning**.

AN INSPIRATION

Hosting of applications on servers within enterprise's data centers



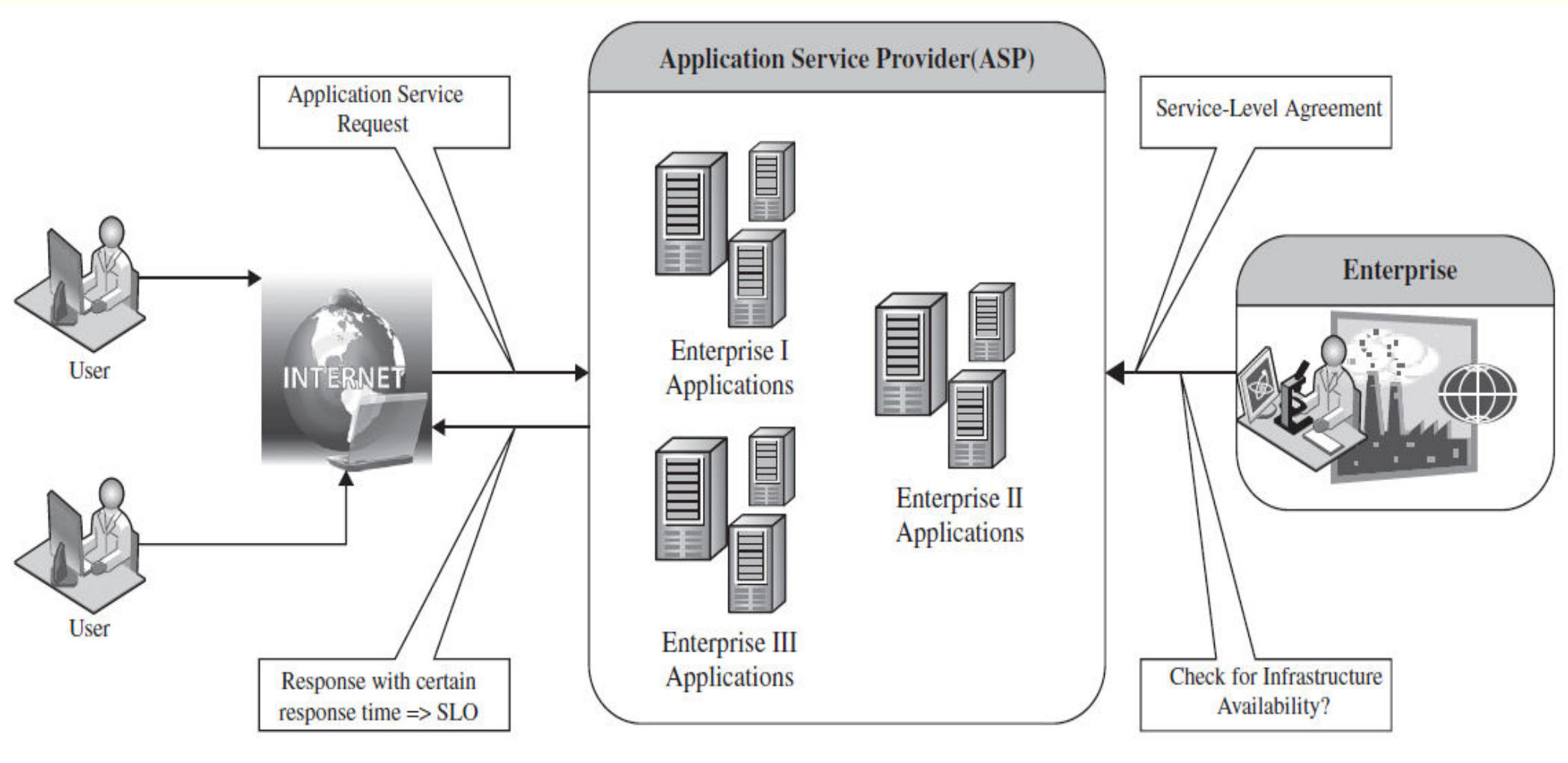
AN INSPIRATION

- Enterprises realized that it was economical to outsource the application hosting activity to third-party infrastructure providers because:
- The enterprises need not invest in procuring expensive hardware upfront without knowing the viability of the business.
- The hardware and application maintenance were non-core activities of their business.
- As the number of web applications grew, the level of sophistication required to manage the data centers increased manyfold—hence the cost of maintaining them.
- The enterprises to enter into a legal agreement with the infrastructure service providers to guarantee a minimum quality of service (QoS).
- The QoS parameters are related to the availability of the system CPU, data storage, and network for efficient execution of the application at peak loads.
- This legal agreement is known as the service-level agreement (SLA).

AN INSPIRATION

- One SLA may state that the application's server machine – 99.9% - core time.
- 85% of the non-core time.
- Reported issue in less than 10 minutes during the core time.
- One hour during non-core time.
- SLAs are known as the infrastructure SLAs, and the infrastructure service providers are known as Application Service Providers (ASPs).

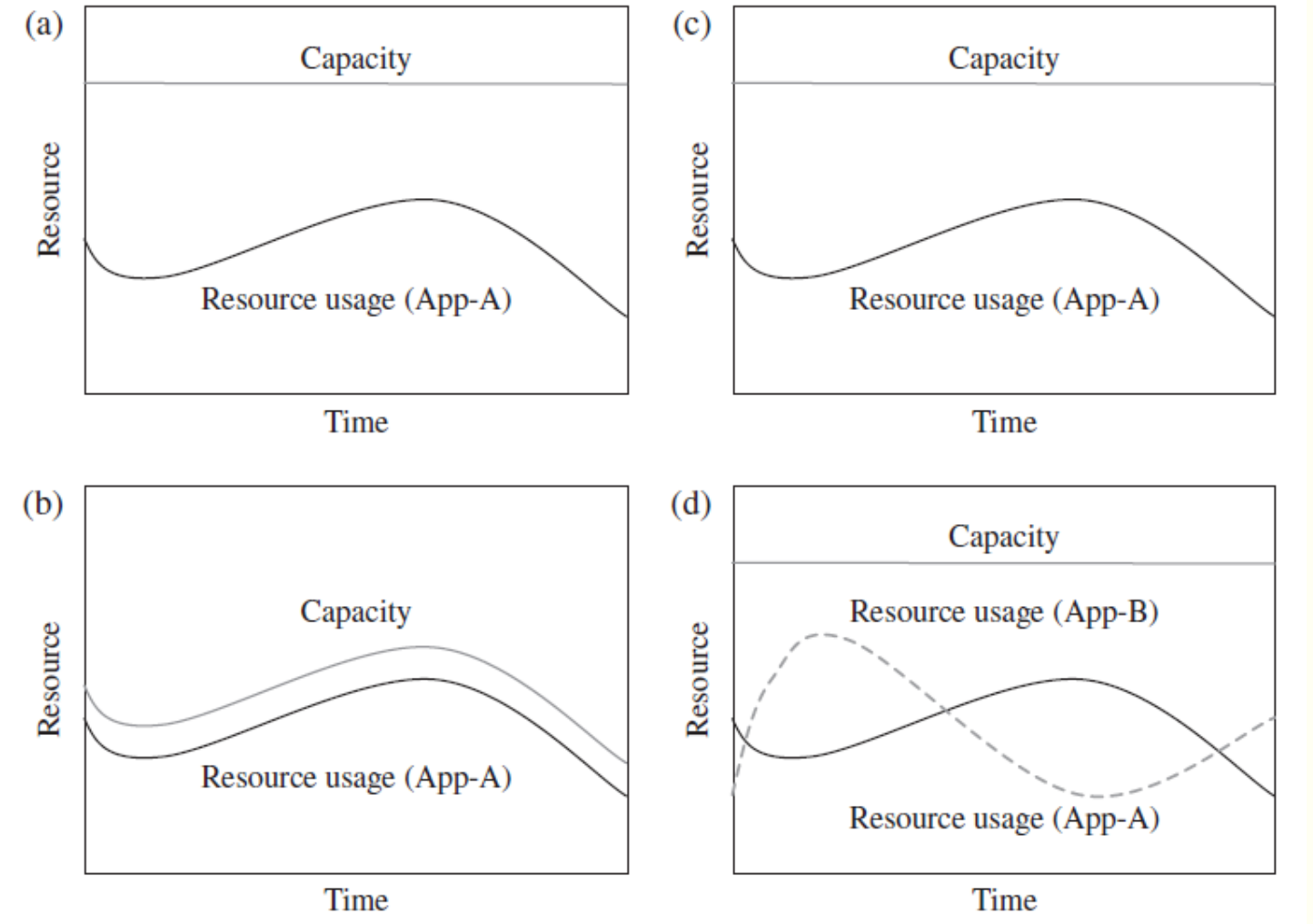
AN INSPIRATION



Dedicated hosting of applications in third party data centers

AN INSPIRATION

Service consumer and service provider perspective before and after the MSP's hosting platforms are virtualized and cloud-enabled. (a) Service consumer perspective earlier. (b) Service consumer perspective now. (c) Service provider perspective earlier. (d) Service provider perspective now.



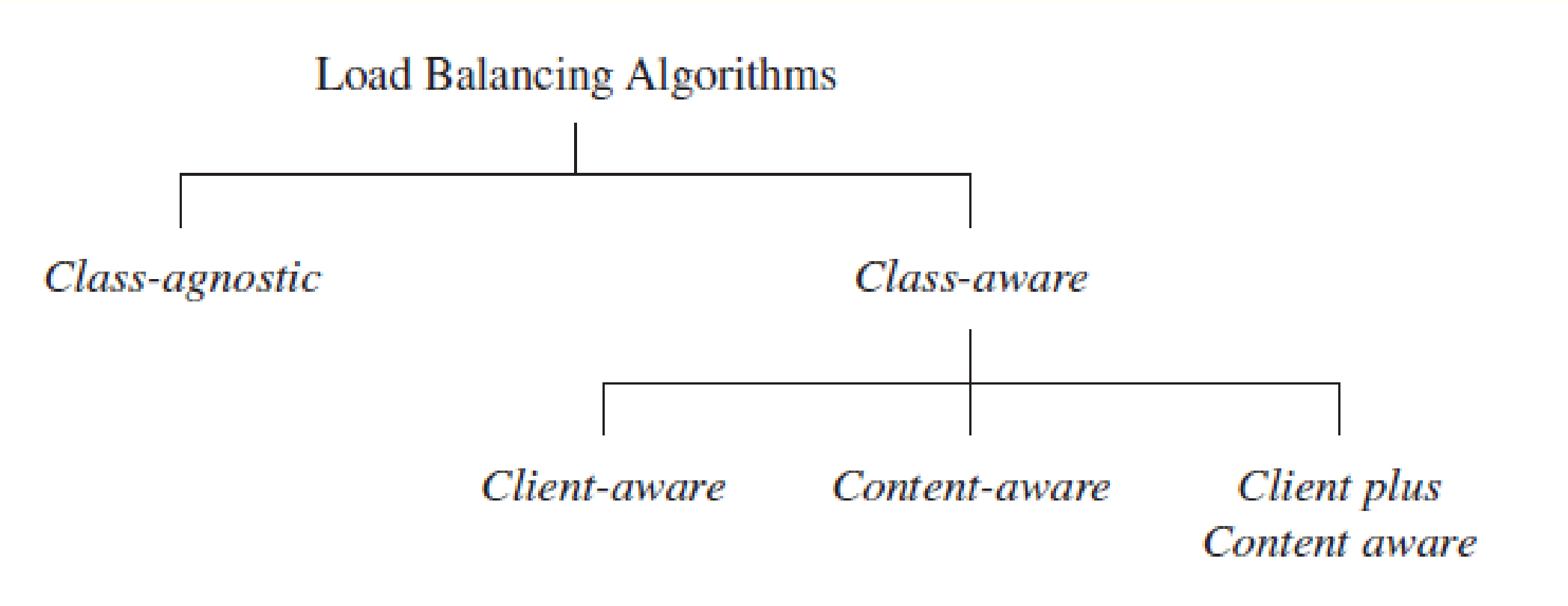
TRADITIONAL APPROACHES TO SLO MANAGEMENT

- Traditionally, load balancing techniques and admission control mechanisms have been used to provide guaranteed quality of service (QoS) for hosted web applications.
- These mechanisms can be viewed as the first attempt towards managing the SLOs.
- In the following subsections we discuss the existing approaches for load balancing and admission control for ensuring QoS.
- Load Balancing
- Admission Control

TRADITIONAL APPROACHES TO SLO MANAGEMENT – LOAD BALANCING

- **Load Balancing**
- The objective of a load balancing is to distribute the incoming requests onto a set of physical machines, each hosting a replica of an application, so that the load on the machines is equally distributed.
- The load balancing algorithm executes on a physical machine that interfaces with the clients.
- This physical machine, also called the front-end node, receives the incoming requests and distributes these requests to different physical machines for further execution.

TRADITIONAL APPROACHES TO SLO MANAGEMENT – LOAD BALANCING

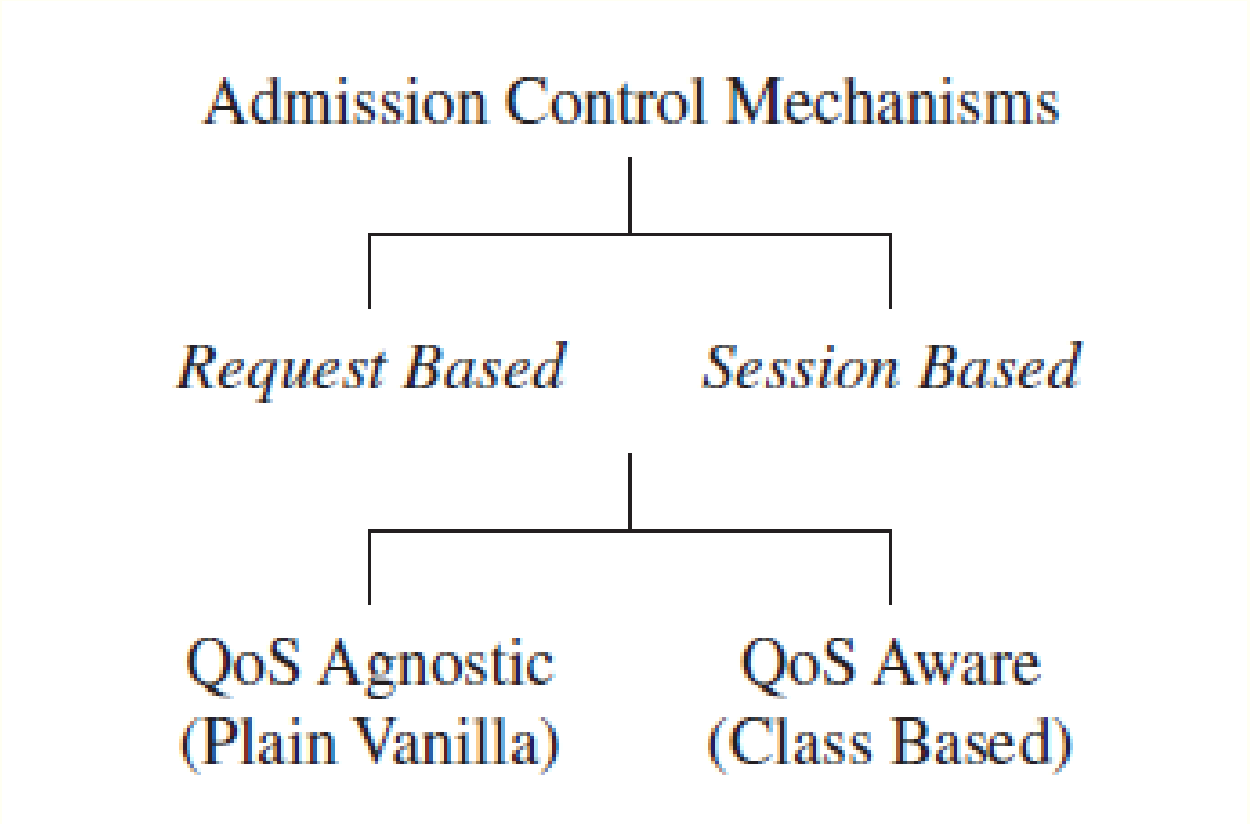


General taxonomy of load-balancing algorithms

TRADITIONAL APPROACHES TO SLO MANAGEMENT - ADMISSION CONTROL

- **Admission Control**
- Admission control algorithms play an important role in deciding the set of requests that should be admitted into the application server when the server experiences “very” heavy loads.
- During overload situations, since the response time for all the requests would invariably degrade if all the arriving requests are admitted into the server, it would be preferable to be selective in identifying a subset of requests that should be admitted into the system so that the overall pay-off is high.
- The objective of admission control mechanisms, therefore, is to police the incoming requests and identify a subset of incoming requests that can be admitted into the system when the system faces overload situations.

TRADITIONAL APPROACHES TO SLO MANAGEMENT - ADMISSION CONTROL



General taxonomy for admission control mechanisms

TYPES OF SLA

- Service-level agreement provides a framework within which both seller and buyer of a service can pursue a profitable service business relationship.
- It outlines the broad understanding between the service provider and the service consumer for conducting business and forms the basis for maintaining a mutually beneficial relationship.
- From a legal perspective, the necessary terms and conditions that bind the service provider to provide services continually to the service consumer are formally defined in SLA.
- SLA can be modeled using web service-level agreement (WSLA) language specification.
- Although WSLA is intended for web-service-based applications, it is equally applicable for hosting of applications.

TYPES OF SLA

TABLE 16.1. Key Components of a Service-Level Agreement

Service-Level Parameter	Describes an observable property of a service whose value is measurable.
Metrics	These are definitions of values of service properties that are measured from a service-providing system or computed from other metrics and constants. Metrics are the key instrument to describe exactly what SLA parameters mean by specifying how to measure or compute the parameter values.
Function	A function specifies how to compute a metric's value from the values of other metrics and constants. Functions are central to describing exactly how SLA parameters are computed from resource metrics.
Measurement directives	These specify how to measure a metric.

TYPES OF SLA

- Infrastructure SLA
- Application SLA

TYPES OF SLA

- **Infrastructure SLA**
- The infrastructure provider manages and offers guarantees on availability of the infrastructure, namely, server machine, power, network connectivity, and so on.
- Enterprises manage themselves, their applications that are deployed on these server machines.
- The machines are leased to the customers and are isolated from machines of other customers.

TYPES OF SLA - INFRASTRUCTURE SLA

TABLE 16.2. Key Contractual Elements of an Infrastructural SLA

Hardware availability	• 99% uptime in a calendar month
Power availability	• 99.99% of the time in a calendar month
Data center network availability	• 99.99% of the time in a calendar month
Backbone network availability	• 99.999% of the time in a calendar month
Service credit for unavailability	• Refund of service credit prorated on downtime period
Outage notification guarantee	• Notification of customer within 1 hr of complete downtime
Internet latency guarantee	• When latency is measured at 5-min intervals to an upstream provider, the average doesn't exceed 60 msec
Packet loss guarantee	• Shall not exceed 1% in a calendar month

TYPES OF SLA

- **Application SLA**
- The application co-location hosting model, the server capacity is available to the applications based solely on their resource demands.
- Hence, the service providers are flexible in allocating and de-allocating computing resources among the co-located applications.
- Therefore, the service providers are also responsible for ensuring to meet their customer's application SLOs.

TYPES OF SLA - APPLICATION SLA

TABLE 16.3. Key contractual components of an application SLA

Service-level parameter metric	• Web site response time (e.g., max of 3.5 sec per user request)
	• Latency of web server (WS) (e.g., max of 0.2 sec per request)
	• Latency of DB (e.g., max of 0.5 sec per query)
Function	• Average latency of WS = (latency of web server 1 + latency of web server 2) /2
	• Web site response time = Average latency of web server + latency of database
Measurement directive	• DB latency available via http://mgmtserver/em/latency
	• WS latency available via http://mgmtserver/ws/instanceno/ latency
Service-level objective	• Service assurance
	• web site latency < 1 sec when concurrent connection < 1000
Penalty	• 1000 USD for every minute while the SLO was breached

LIFE CYCLE OF SLA

- Each SLA goes through a sequence of steps starting from identification of terms and conditions, activation and monitoring of the stated terms and conditions, and eventual termination of contract once the hosting relationship ceases to exist.
- Such a sequence of steps is called SLA life cycle and consists of the following five phases:

1. Contract definition
2. Publishing and discovery
3. Negotiation
4. Operationalization
5. De-commissioning

LIFE CYCLE OF SLA

- **Contract Definition**
- Generally, service providers define a set of service offerings and corresponding SLAs using standard templates.
- These service offerings form a catalog.
- Individual SLAs for enterprises can be derived by customizing these base SLA templates.

LIFE CYCLE OF SLA

- **Publication and Discovery**
- Service provider advertises these base service offerings through standard publication media, and the customers should be able to locate the service provider by searching the catalog.
- The customers can search different competitive offerings and shortlist a few that fulfill their requirements for further negotiation.

LIFE CYCLE OF SLA

- **Negotiation**
- Once the customer has discovered a service provider who can meet their application hosting need, the SLA terms and conditions needs to be mutually agreed upon before signing the agreement for hosting the application.
- For a standard packaged application which is offered as service, this phase could be automated.
- For customized applications that are hosted on cloud platforms, this phase is manual.
- The service provider needs to analyze the application's behavior with respect to scalability and performance before agreeing on the specification of SLA.
- At the end of this phase, the SLA is mutually agreed by both customer and provider and is eventually signed off.

LIFE CYCLE OF SLA

- **Operationalization**
- SLA operation consists of SLA monitoring, SLA accounting, and SLA enforcement.
- SLA monitoring involves measuring parameter values and calculating the metrics defined as a part of SLA and determining the deviations.
- On identifying the deviations, the concerned parties are notified.
- SLA accounting involves capturing and archiving the SLA adherence for compliance.

LIFE CYCLE OF SLA

- **De-commissioning**
- SLA decommissioning involves termination of all activities performed under a particular SLA when the hosting relationship between the service provider and the service consumer has ended.
- SLA specifies the terms and conditions of contract termination and specifies situations under which the relationship between a service provider and a service consumer can be considered to be legally ended.

SLA MANAGEMENT IN CLOUD

- SLA management of applications hosted on cloud platforms involves five phases.
  1. Feasibility
  2. On-boarding
  3. Pre-production
  4. Production
  5. Termination

SLA MANAGEMENT IN CLOUD

- **Feasibility Analysis**

- MSP conducts the feasibility study of hosting an application on their cloud platforms. This study involves three kinds of feasibility: (1) technical feasibility, (2) infrastructure feasibility, and (3) financial feasibility. The technical feasibility of an application implies determining the following:

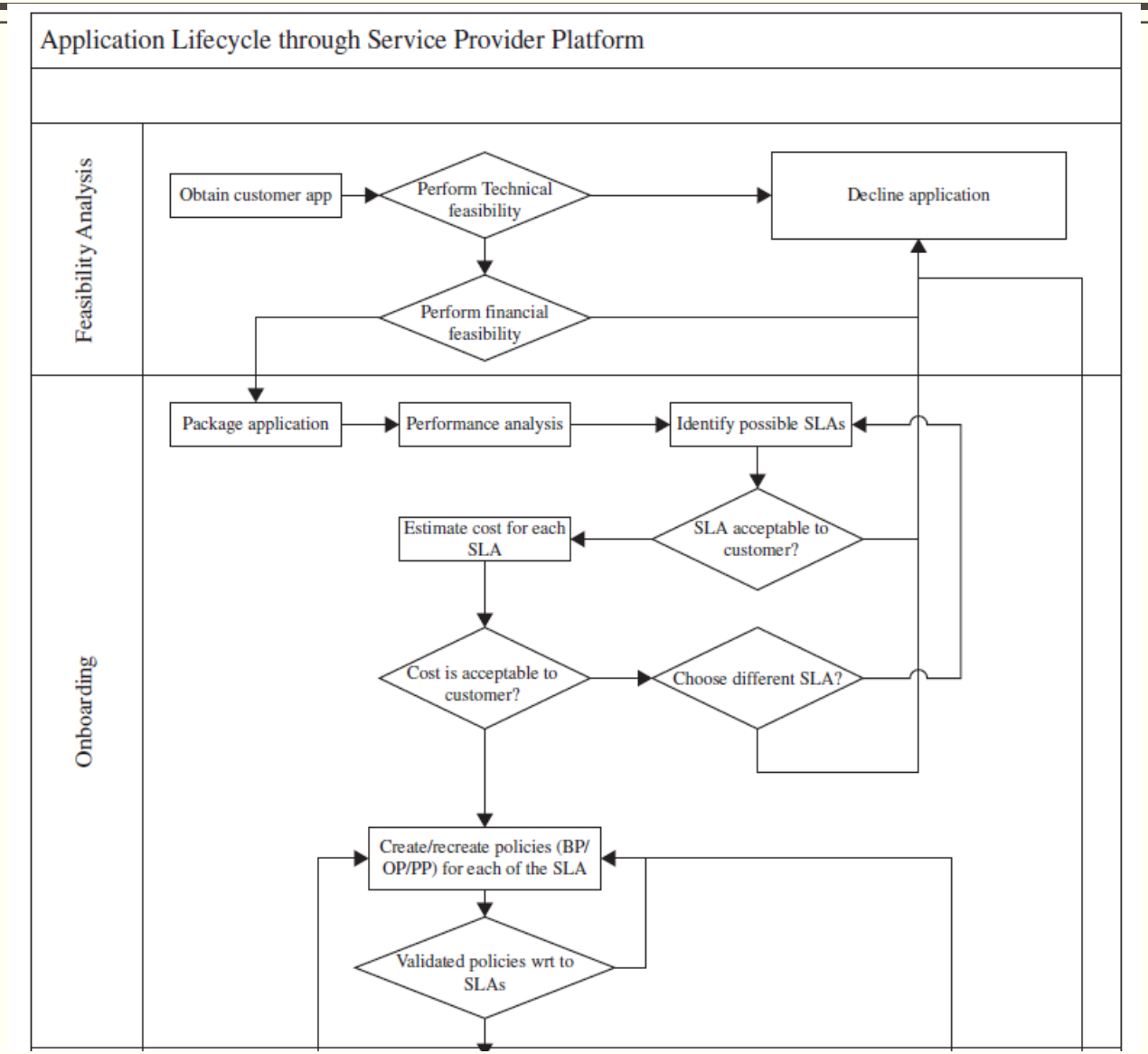
  1. Ability of an application to scale out.

  2. Compatibility of the application with the cloud platform being used within the MSP's data center.

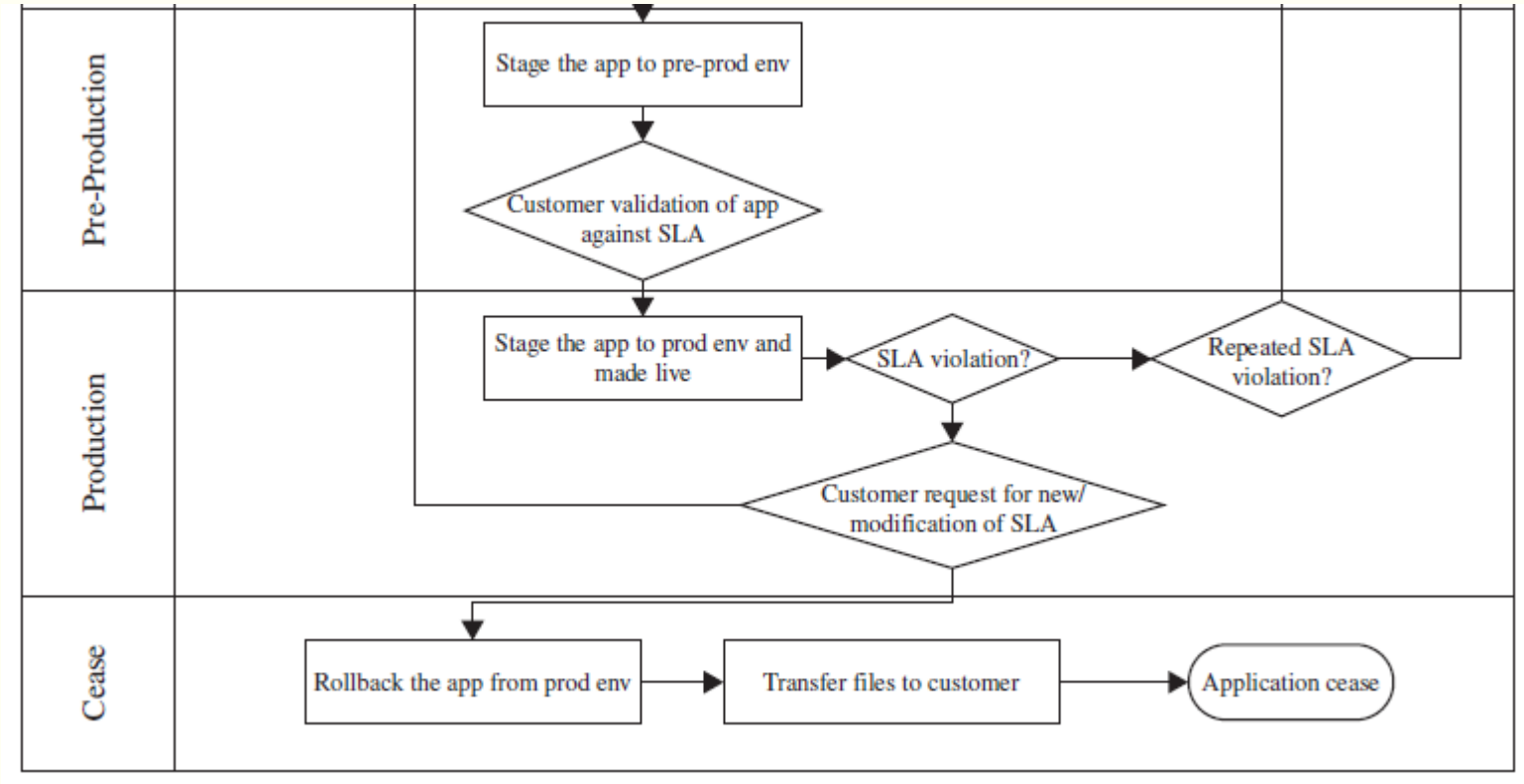
  3. The need and availability of a specific hardware and software required for hosting and running of the application.

  4. Preliminary information about the application performance and whether they can be met by the MSP.

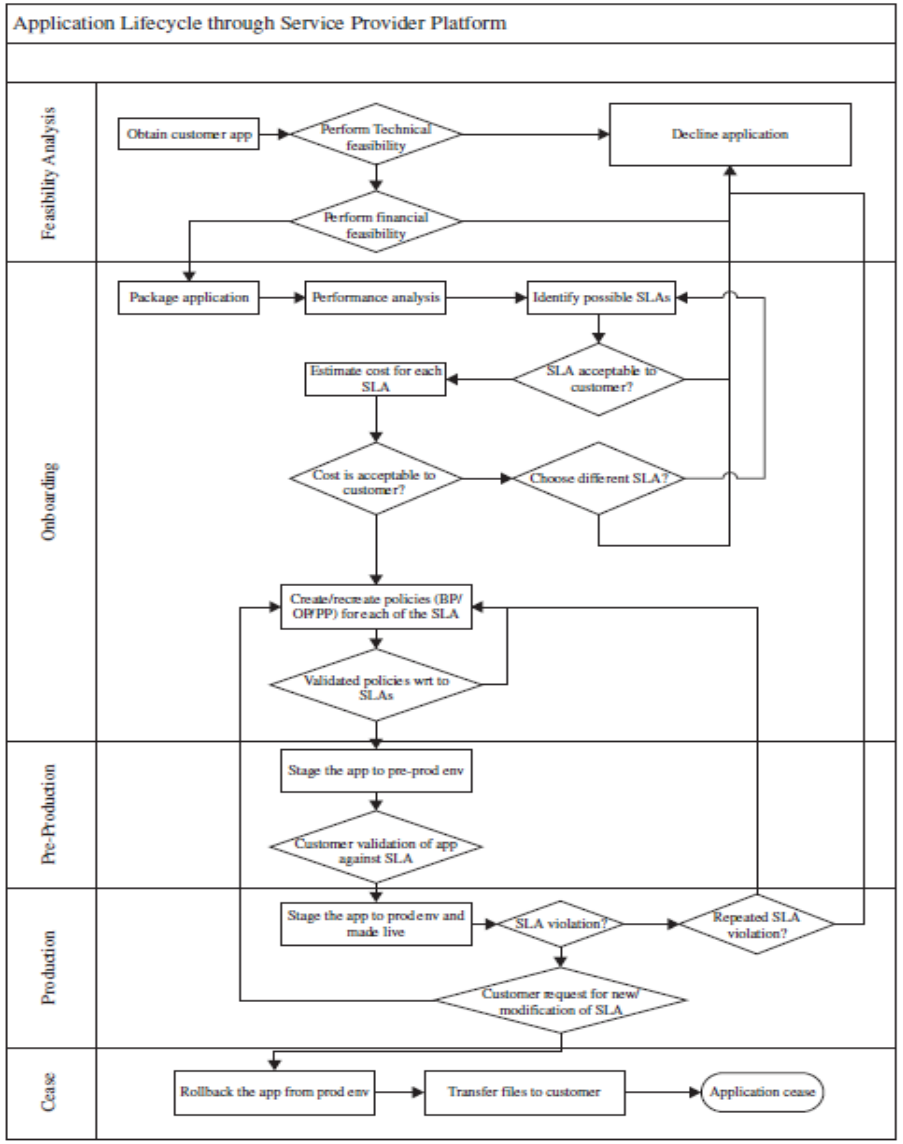
SLA MANAGEMENT IN CLOUD



SLA MANAGEMENT IN CLOUD



SLA MANAGEMENT IN CLOUD



SLA MANAGEMENT IN CLOUD

- **On-Boarding of Application**
- Once the customer and the MSP agree in principle to host the application based on the findings of the feasibility study, the application is moved from the customer servers to the hosting platform.
- Moving an application to the MSP's hosting platform is called on-boarding.
- As part of the on-boarding activity, the MSP understands the application runtime characteristics using runtime profilers.
- This helps the MSP to identify the possible SLAs that can be offered to the customer for that application.
- This also helps in creation of the necessary policies (also called rule sets) required to guarantee the SLOs mentioned in the application SLA. The application is accessible to its end users only after the on-boarding activity is completed.

SLA MANAGEMENT IN CLOUD

- **Preproduction**
- Once the determination of policies is completed as discussed in previous phase, the application is hosted in a simulated production environment.
- It facilitates the customer to verify and validate the MSP's findings on application's runtime characteristics and agree on the defined SLA.
- Once both parties agree on the cost and the terms and conditions of the SLA, the customer sign-off is obtained.
- On successful completion of this phase the MSP allows the application to go on-live.

SLA MANAGEMENT IN CLOUD

- **Production**
- In this phase, the application is made accessible to its end users under the agreed SLA.
- However, there could be situations when the managed application tends to behave differently in a production environment compared to the preproduction environment.
- This in turn may cause sustained breach of the terms and conditions mentioned in the SLA.

SLA MANAGEMENT IN CLOUD

- **Production**
- Additionally, customer may request the MSP for inclusion of new terms and conditions in the SLA.
- If the application SLA is breached frequently or if the customer requests for a new non-agreed SLA, the on-boarding process is performed again.
- In the case of the former, on-boarding activity is repeated to analyse the application and its policies with respect to SLA fulfillment.
- In case of the latter, a new set of policies are formulated to meet the fresh terms and conditions of the SLA.

SLA MANAGEMENT IN CLOUD

- **Termination**
- When the customer wishes to withdraw the hosted application and does not wish to continue to avail the services of the MSP for managing the hosting of its application, the termination activity is initiated.
- On initiation of termination, all data related to the application are transferred to the customer and only the essential information is retained for legal compliance.
- This ends the hosting relationship between the two parties for that application, and the customer sign-off is obtained.

AUTOMATED POLICY-BASED MANAGEMENT

- This section explains in detail the operationalization of the “Operational” and “Provisioning” policies defined as part of the on-boarding activity.
- The policies specify the sequence of actions to be performed under different circumstances.
- Operational policies specify the functional relationship between the system level infrastructural attributes and the business level SLA goals.
- Knowledge of such a relationship helps in identifying the quantum of system resources to be allocated to the various components of the application for different system attributes at various workloads, workload compositions, and operating conditions, so that the SLA goals are met.

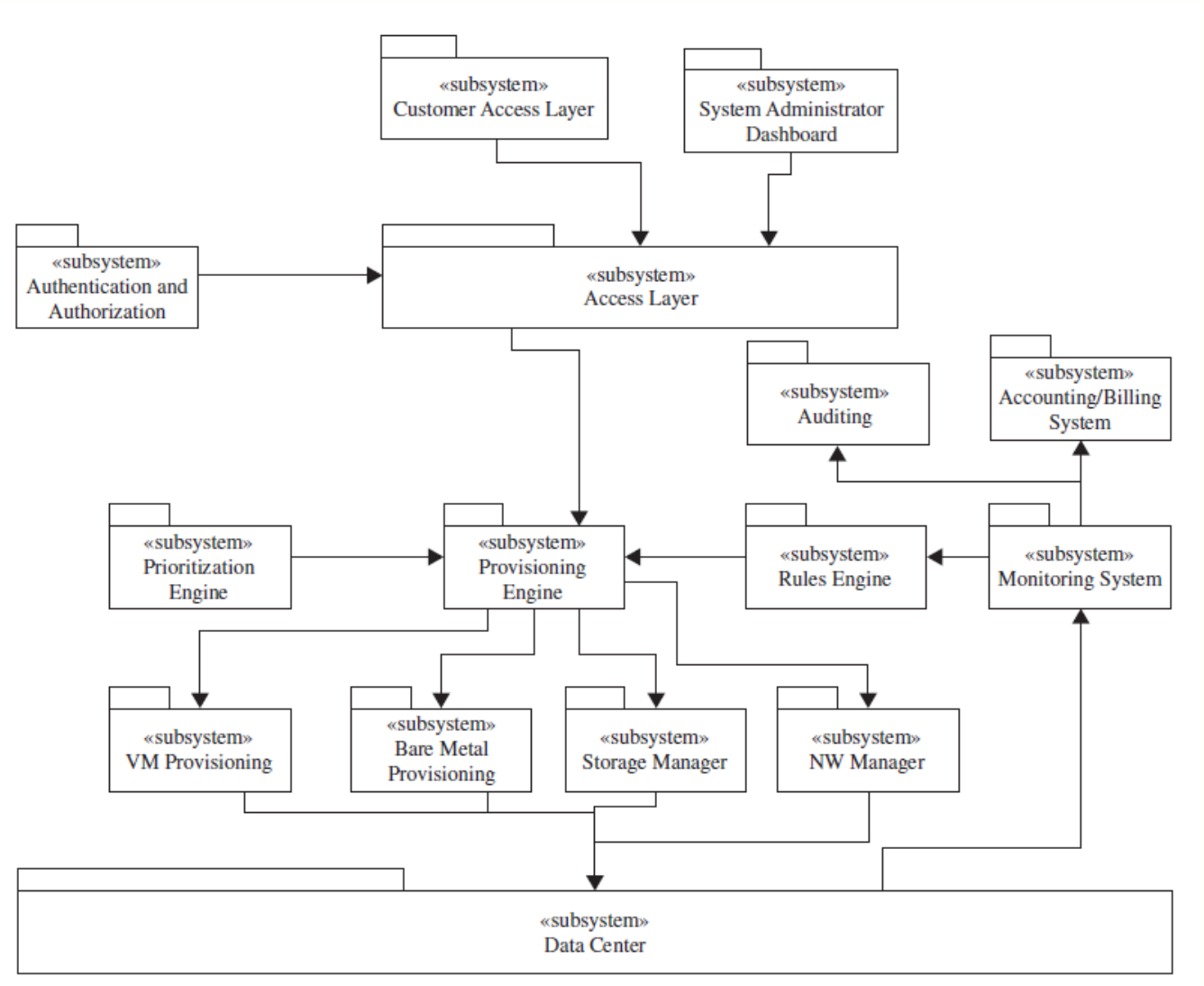
AUTOMATED POLICY-BASED MANAGEMENT

- This section explains in detail the operationalization of the “Operational” and “Provisioning” policies defined as part of the on-boarding activity.
- The policies specify the sequence of actions to be performed under different circumstances.
- Operational policies specify the functional relationship between the system level infrastructural attributes and the business level SLA goals.
- Knowledge of such a relationship helps in identifying the quantum of system resources to be allocated to the various components of the application for different system attributes at various workloads, workload compositions, and operating conditions, so that the SLA goals are met.

AUTOMATED POLICY-BASED MANAGEMENT

- The parameters often used to prioritize action and perform resource contention resolution are:
  - The SLA class (Platinum, Gold, Silver, etc.) to which the application belongs to.
  - The amount of penalty associated with SLA breach.
  - Whether the application is at the threshold of breaching the SLA.
  - Whether the application has already breached the SLA.
  - The number of applications belonging to the same customer that has breached SLA.
  - The number of applications belonging to the same customer about to breach SLA.
  - The type of action to be performed to rectify the situation.

AUTOMATED POLICY-BASED MANAGEMENT



AUTOMATED POLICY-BASED MANAGEMENT

- The basic functionality of these components is described below:

1. **Prioritization Engine.** Requests from different customers' web applications contending for the same resource are identified, and accordingly their execution is prioritized. Business policies defined by the MSP helps in identifying the requests whose execution should be prioritized in case of resource contentions so that the MSP can realize higher benefits.

2. **Provisioning Engine.** Every user request of an application will be enacted by the system. The set of steps necessary to enact the user requests are defined in the provisioning policy, and they are used to fulfill the application request like starting an application, stopping an application, and so on. These set of steps can be visualized as a workflow. Hence, the execution of provisioning policy requires a workflow engine.

AUTOMATED POLICY-BASED MANAGEMENT

3. **Rules Engine.** The operation policy defines a sequence of actions to be enacted under different conditions/trigger points. The rules engine evaluates the data captured by the monitoring system, evaluates against the predefined operation rules, and triggers the associated action if required. Rules engine and the operational policy is the key to guaranteeing SLA under a self healing system.

4. **Monitoring System.** Monitoring system collects the defined metrics in SLA. These metrics are used for monitoring resource failures, evaluating operational policies, and auditing and billing purpose.

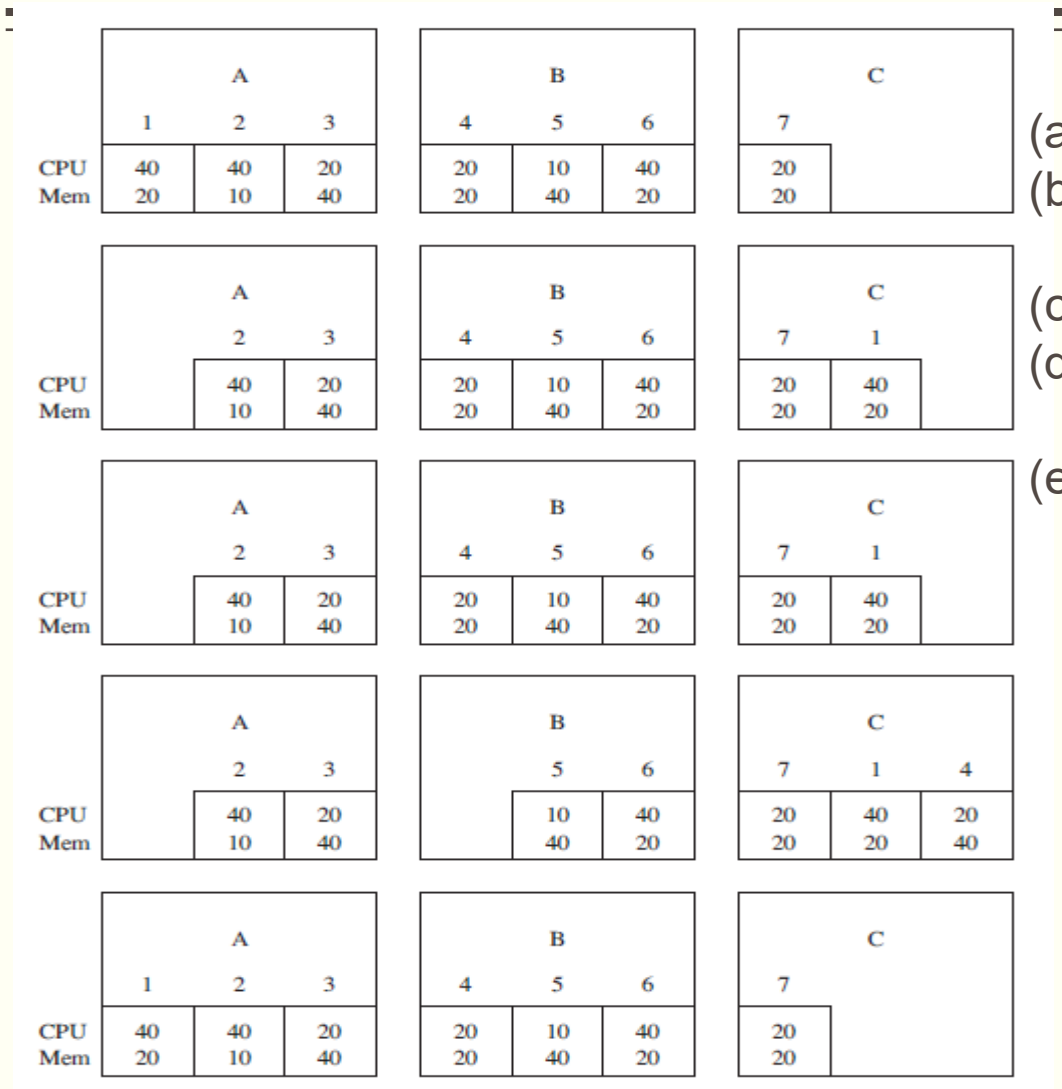
5. **Auditing.** The adherence to the predefined SLA needs to be monitored and recorded. It is essential to monitor the compliance of SLA because any noncompliance leads to strict penalties. The audit report forms the basis for strategizing and long-term planning for the MSP.

6. **Accounting/Billing System.** Based on the payment model, chargebacks could be made based on the resource utilized by the process during the operation. The fixed cost and recurring costs are computed and billed accordingly.

AUTOMATED POLICY-BASED MANAGEMENT

- Assume an initial assignment of seven virtual machines (VM) to the three physical machines (PM) at time t1. Also, each of the three PMs has memory and CPU capacity of 100. At time t1, the CPU usage by VM1, VM2, and VM3 on PMA are 40, 40, and 20, respectively, and the memory consumption is 20, 10, and 40 respectively. Similarly, at time t1 the CPU and memory requirements of VM4, VM5, and VM6 on PMB are 20, 10, 40 and 20, 40, 20, respectively. VM7 only consumes 20% of CPU and 20% of memory on PMC. Thus, PMB and PMC are underloaded but PMA is overloaded. Assume VM1 is the cause of the overload situation in PMA.

AUTOMATED POLICY-BASED MANAGEMENT



(a) Initial configuration of the VMs and the PMs at time t_1.
(b) Configuration resulting from event-based migration of VM_1 at time t_1.
(c) Resource requirement situation at time t_2 . t_1.
(d) Configuration resulting from “event-based” migration of VM_4 at time t_2 . t_1.
(e) Alternate configuration resulting from optimization based migration at time t_2 . t_1.

CLOUD MANAGEMENT STANDARDS

- Cloud computing management is maintaining and controlling the cloud services and resources be it public, private or hybrid.
- Some of its aspects include load balancing, performance, storage, backups, capacity, deployment etc.
- To do so a cloud managing personnel needs full access to all the functionality of resources in the cloud.
- Different software products and technologies are combined to provide a cohesive cloud management strategy and process.

CLOUD MANAGEMENT STANDARDS

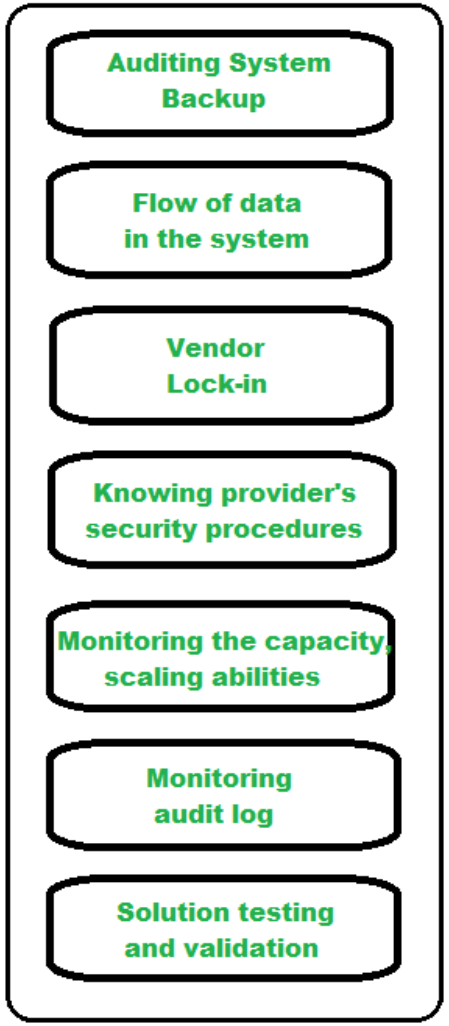
- **Need of Cloud Management :**
- Cloud is nowadays preferred by huge organizations as their primary data storage.
- A small downtime or an error can cause a great deal of loss and inconvenience for the organizations.
- So as to design, handle and maintain a cloud computing service specific members are responsible who make sure things work out as supposed and all arising issues are addressed.

CLOUD MANAGEMENT STANDARDS

- **Cloud Management Platform :**
- A cloud management platform is a software solution that has a robust and extensive set of APIs that allow it to pull data from every corner of the IT infrastructure.
- A CMP allows an IT organization to establish a structured approach to security and IT governance that can be implemented across the organization’s entire cloud environment.

CLOUD MANAGEMENT STANDARDS

- **Cloud Management Tasks :**



CLOUD MANAGEMENT STANDARDS

- **Cloud Management Tasks :**
- **Auditing System Backups –** It is required to audit the backups from time to time to ensure restoration of randomly selected files of different users. This might be done by the organization or by the cloud provider.
- **Flow of data in the system –** The managers are responsible for designing a data flow diagram that shows how the data is supposed to flow throughout the organization.
- **Vendor Lock-In –** The managers should know how to move their data from a server to another in case the organization decides to switch providers.
- **Knowing provider's security procedures –** The managers should know the security plans of the provider, especially Multitenant use, E-commerce processing, Employee screening and Encryption policy.

CLOUD MANAGEMENT STANDARDS

- **Cloud Management Tasks :**
- **Monitoring the Capacity, Planning and Scaling abilities –** The manager should know if their current cloud provider is going to meet their organization's demand in the future and also their scaling capabilities.
- **Monitoring audit log –** In order to identify errors in the system, logs are audited by the managers on a regular basis.
- **Solution Testing and Validation –** It is necessary to test the cloud services and verify the results and for error-free solutions.

REFERENCES

1. Kai Hwang, Geoffrey C Fox and Jack G Dongarra, "Distributed and Cloud Computing, From Parallel Processing to the Internet of Things", Morgan Kaufmann Publishers, 2012.
2. Barrie Sosinky,"Cloud Computing Bible", Wiley Publishing Inc,2011
3. Buyya R., Broberg J. and Goscinski A., "Cloud Computing: Principles and Paradigm", First Edition, John Wiley & Sons, 2011.
4. Rajkumar Buyya, Christian Vecchiola, S. ThamaraiSelvi,"Mastering the Cloud Computing", Morgan Kaufmann,2013
5. John W. Rittinghouse and James F. Ransome, "Cloud Computing: Implementation "Management, and Security", CRC Press, 2016.
6. David Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes", IEEE Cloud Computing, Volume: 1 , Issue: 3 , 2014.
7. VMware (white paper),"Understanding Full Virtualization, Paravirtualization, and Hardware Assist ":www.vmware.com/files/pdf/VMware_paravirtualization.pdf.

Thank You…