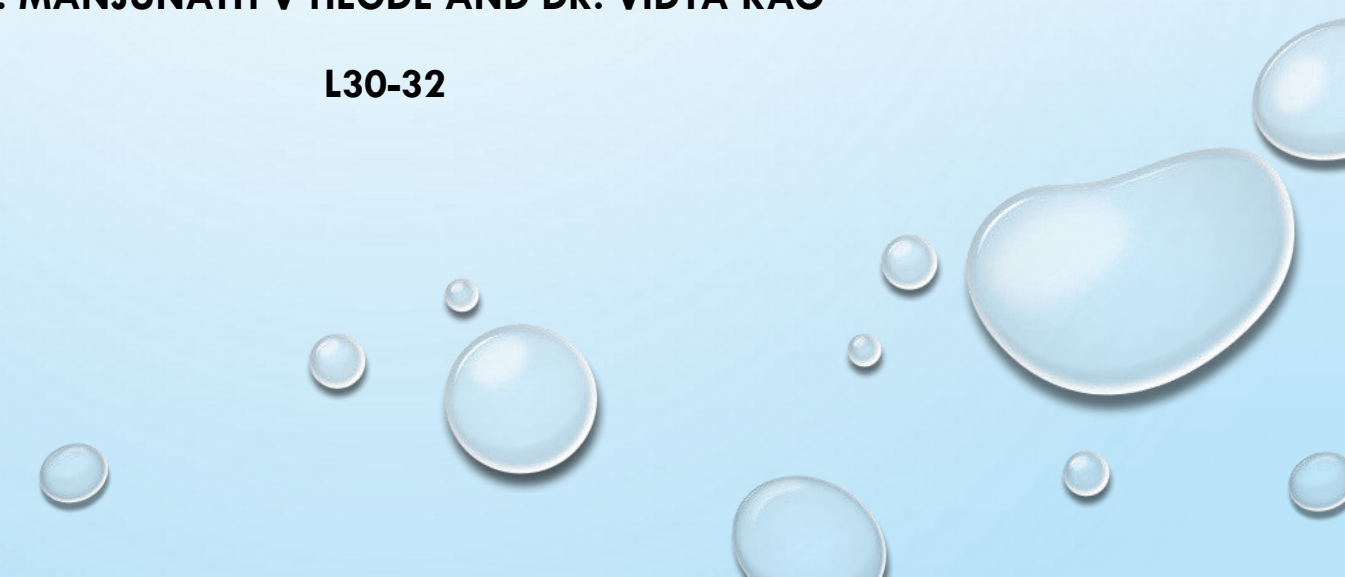




SERVICE LEVEL AGREEMENTS (SLA)

DR. MANJUNATH V HEGDE AND DR. VIDYA RAO

L30-32



INSPIRATION

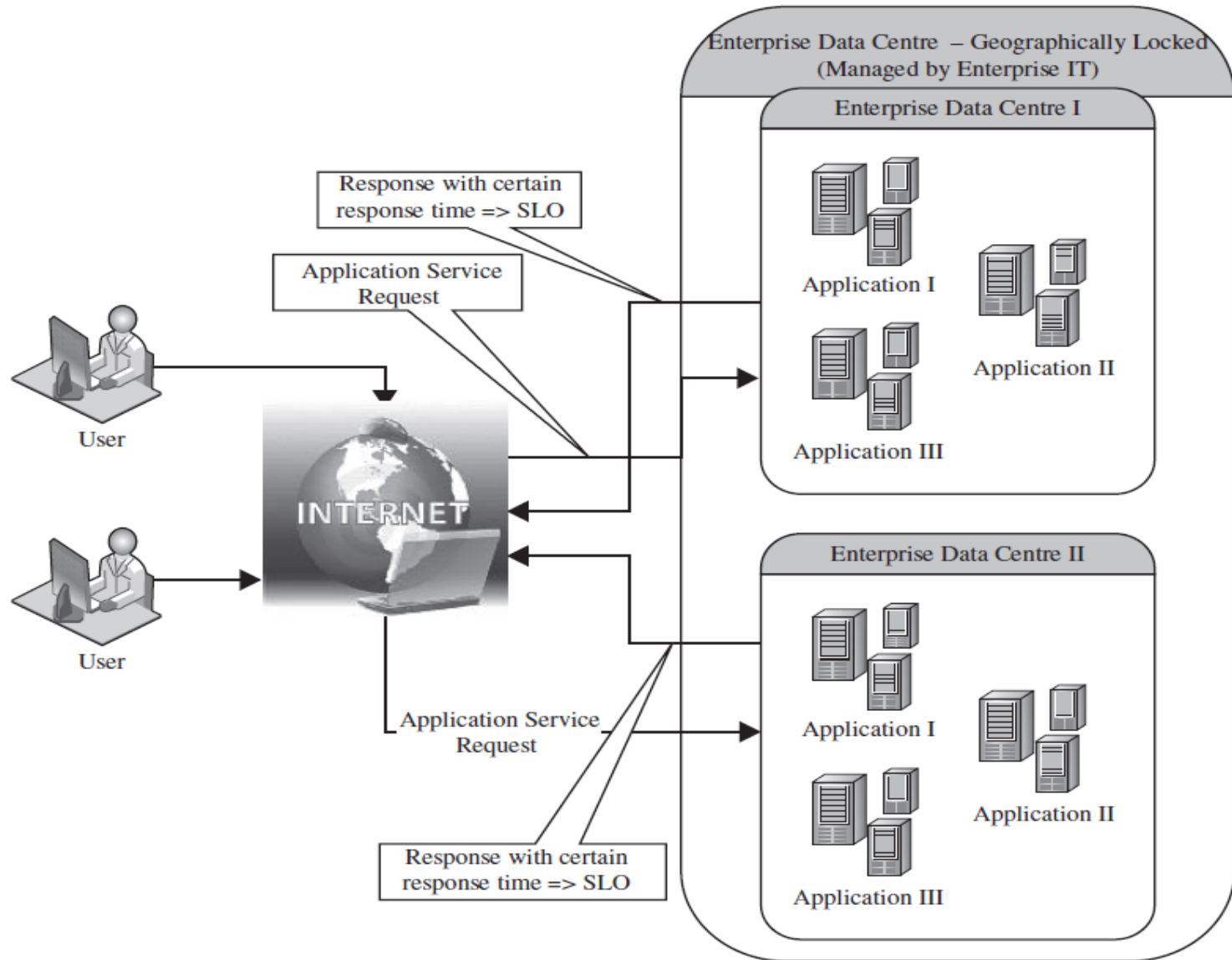


FIGURE 16.1. Hosting of applications on servers within enterprise's data centers.

INSPIRATION (CONTD..)

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.

INSPIRATION (CONTD..)

- Enterprises developed the **web applications and deployed** on the infrastructure of the third-party service providers.
- These providers get **the required hardware** and make it available for application hosting.
- It necessitated the enterprises to enter into a **legal agreement** with the infrastructure service providers to guarantee a minimum quality of service (QoS).
- Typically, 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).**

INSPIRATION (CONTD..)

- Consider an example, one SLA may state that the **application's server machine will be available for 99.9%** of the **key business hours** of the application's end users, also called **core time**, and **85% of the non-core time**.
- Another SLA may state that the service provider would **respond to a reported issue in less than 10 minutes during the core time**, but **would respond in one hour during non-core time**.
- These SLAs are known as the **infrastructure SLAs**, and the infrastructure service providers are known as **Application Service Providers (ASPs)**.

INSPIRATION (CONTD..)

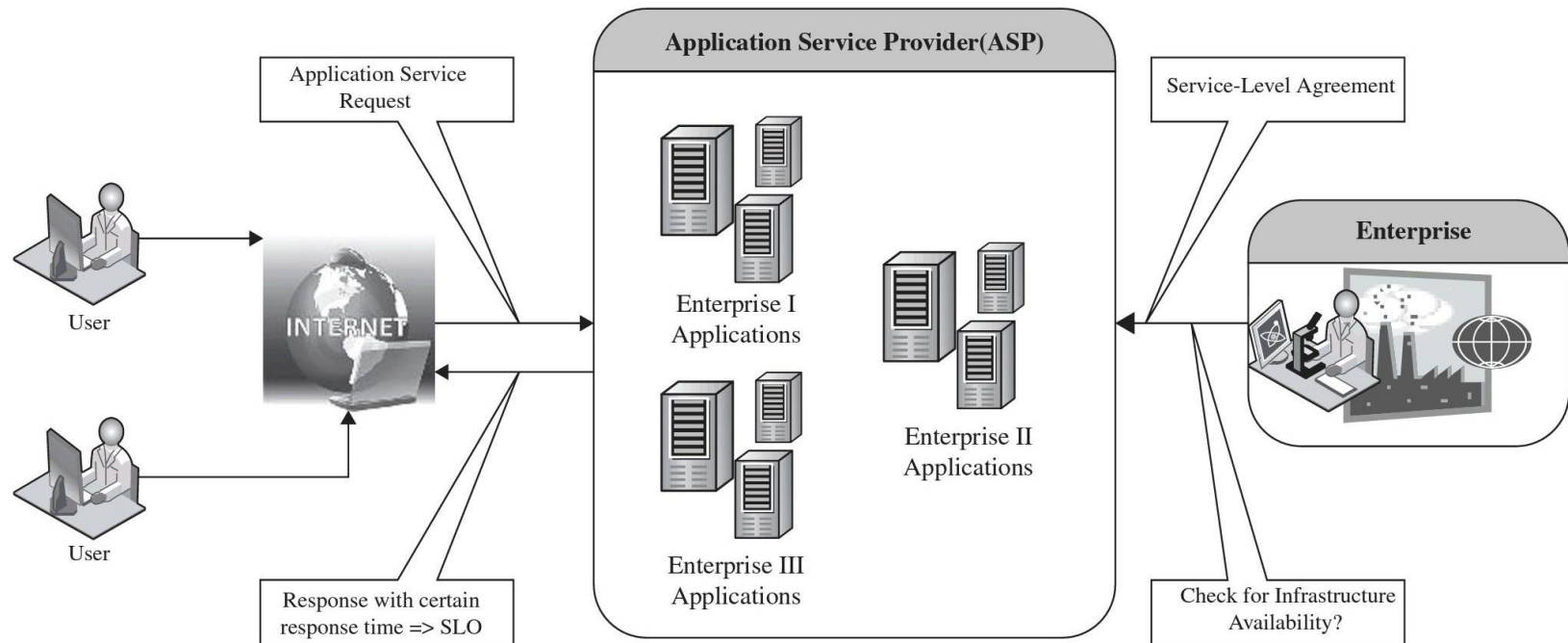


FIGURE 16.2. Dedicated hosting of applications in third party data centers.

INSPIRATION (CONTD..)

- Consequently, a set of tools for **monitoring and measurement** of availability of the infrastructure were required and developed.
- However, availability of the infrastructure doesn't automatically guarantee the availability of the application for its end users.
- These tools helped in tracking the **SLA adherence**.
- The responsibility for making the application available to its end users is with the enterprises.
- Therefore, the enterprises' IT team performs **capacity planning, and the infrastructure** provider procures the same.

INSPIRATION (CONTD..)

- The **dedicated hosting practice resulted in massive redundancies** within the ASP's data centers due to the underutilization of many of their servers.
- This is because the **applications were not fully utilizing their servers'** capacity at **nonpeak loads**.
- To reduce the redundancies and increase the server utilization in data centers, ASPs started **co-hosting applications** with complementary workload patterns.
- **Co-hosting of applications means deploying more than one application on a single server.**

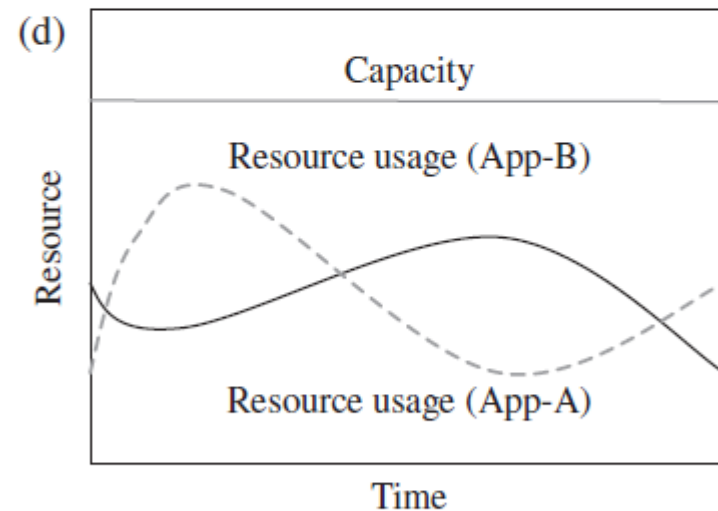
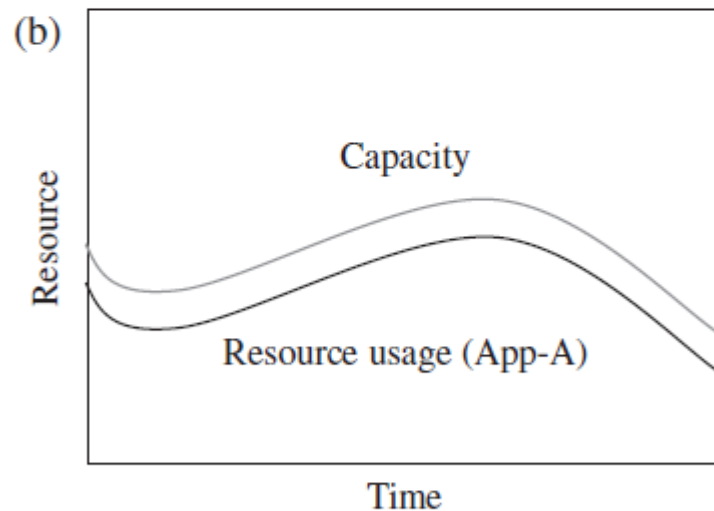
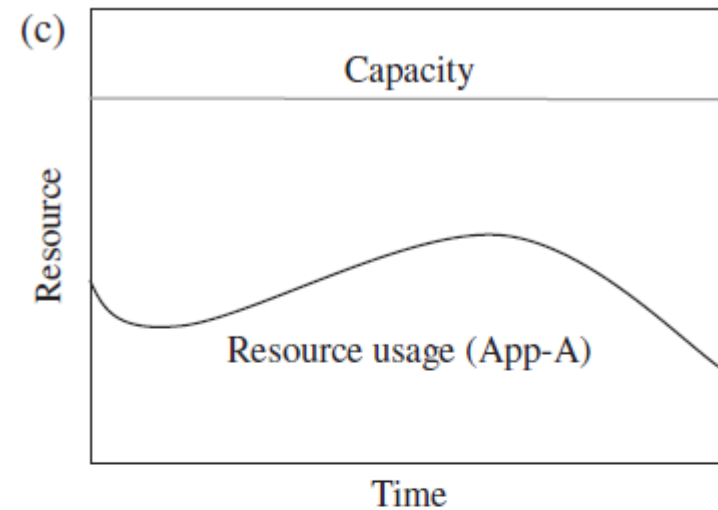
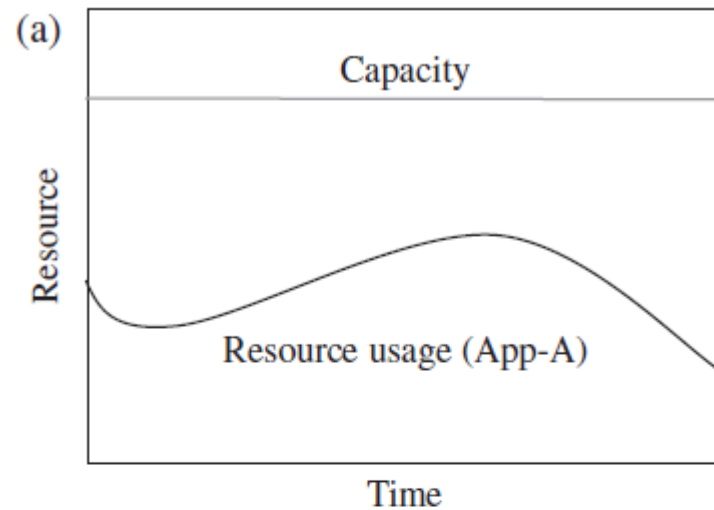


FIGURE 16.3. Service consumer and service provider perspective before and after the MSP's hosting platforms are virtualized and cloud-enabled. (a) Service consumer perspective earlier. (b) Service consumer perspective now. (c) Service provider perspective earlier. (d) Service provider perspective now.

INSPIRATION (CONTD..)

Performance Isolation

Security

Co-Hosting disadvantages

Performance isolation:

- one application **should not steal the resources** being utilized by other co-located applications.
- For example, assume that application **A** is required to use more quantity of a resource than originally allocated to it for duration of time ***t***.
- For that duration the amount of the same resource available to application **B** is decreased.
- This could adversely affect the performance of application **B**.

Security

one application should not access and destroy the data and other information of co-located applications.

To handle the above issues

**Virtualization
technologies**

INSPIRATION (CONTD..)

Virtualization technologies

- The applications, instead of being hosted on the physical machines, can be encapsulated using virtual machines.
- These virtual machines are then mapped to the physical machines.
- System resource allocation to these virtual machines can be made in two modes: **(1) conserving** and **(2) non-conserving**.

A virtual machine demanding more system resources (CPU and memory) than the specified quota cannot be allocated the spare resources that remain unutilized by the other co-hosted virtual machines.

Here, the spare resources that are not utilized by the co-hosted virtual machines can be used by the virtual machine needing the extra amount of resource.

TRADITIONAL APPROACHES TO SLO MANAGEMENT

First attempt to balance QoS among service-level objectives (SLOs)



**Load balancing
techniques**

**Admission
control
mechanisms**

TRADITIONAL APPROACHES TO SLO MANAGEMENT

Load balancing techniques

- 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.
- This set of physical machines is responsible for serving the incoming requests and are known as the back-end nodes.
-

TRADITIONAL APPROACHES TO SLO MANAGEMENT

Load balancing techniques

- Typically, the algorithm executing on the front-end node is agnostic to the nature of the request.
- This means that the front-end node is neither aware of the type of client from which the request originates nor aware of the category to which the request belongs to ----- **class agnostic**
- In **class-aware** load balancing and requests distribution, the front-end node must additionally inspect the type of client making the request and/or the type of service requested before deciding which back-end node should service the request.

TRADITIONAL APPROACHES TO SLO MANAGEMENT

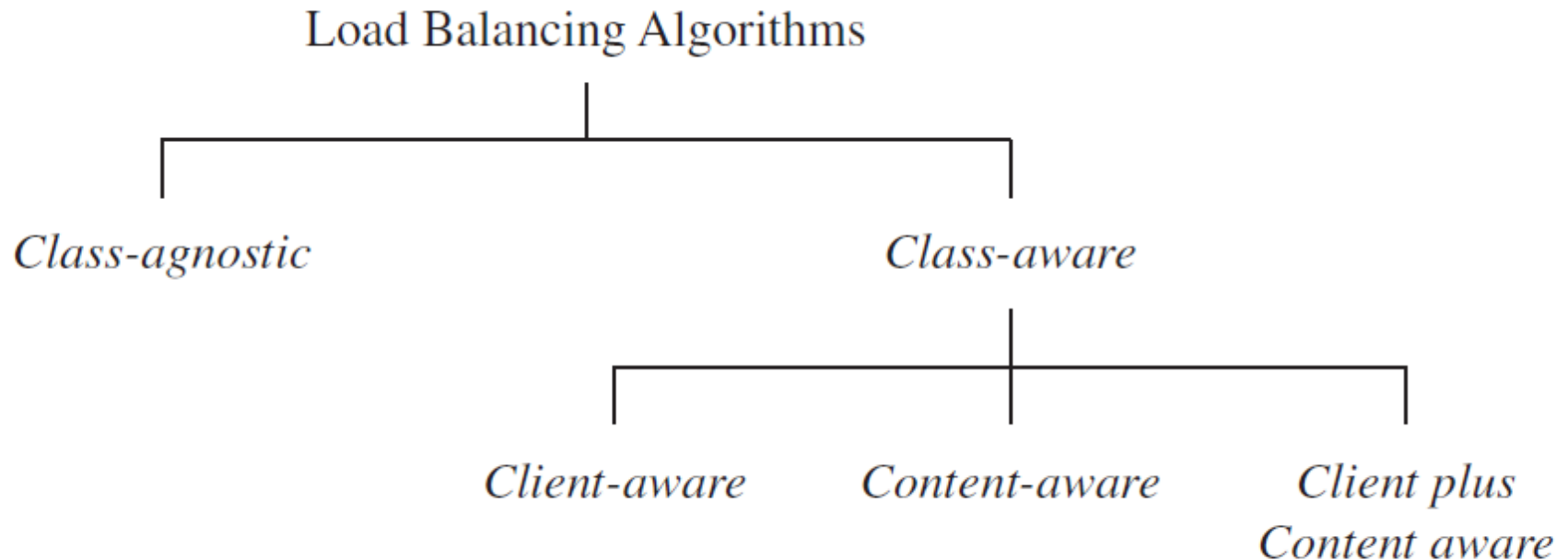


FIGURE 16.5. General taxonomy of load-balancing algorithms.

TRADITIONAL APPROACHES TO SLO MANAGEMENT

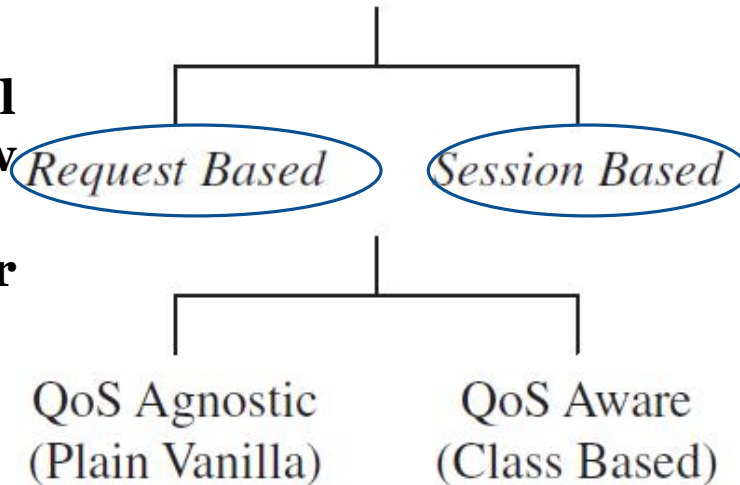
Admission Control

- 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.
- 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

Admission Control Mechanisms



Request-based admission control algorithms reject new requests if the servers are running to their capacity.

Session-based admission control mechanisms try to ensure that longer sessions are completed and any new sessions are rejected.

FIGURE 16.6. General taxonomy for admission control mechanisms.

TYPES OF SLA



```
graph TD; A[TYPES OF SLA] --> B[Infrastructure SLA]; A --> C[Application SLA]; B --> D[• The infrastructure provider manages and offers guarantees on availability of the infrastructure, namely, server machine, power, network connectivity, and so on.]; B --> E[• The machines are leased to the customers and are isolated from machines of other customers]; C --> F[• In the application co-location hosting model, the server capacity is available to the applications based solely on their resource demands.]; C --> G[• The service providers are flexible in allocating and de-allocating computing resources among the co-located applications.];
```

Infrastructure SLA

- **The infrastructure provider manages and offers guarantees on availability of the infrastructure, namely, server machine, power, network connectivity, and so on.**
- **The machines are leased to the customers and are isolated from machines of other customers**

Application SLA

- **In the application co-location hosting model, the server capacity is available to the applications based solely on their resource demands.**
- **The service providers are flexible in allocating and de-allocating computing resources among the co-located applications.**

TYPES OF SLA

TABLE 16.1. Key Components of a Service-Level Agreement

| | |
|---------------------------------------|--|
| Service-Level Parameter Metrics | <p>Describes an observable property of a service whose value is measurable.</p> <p>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.</p> |
| Function | <p>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.</p> |
| Measurement directives | <p>These specify how to measure a metric.</p> |

TYPES OF SLA

TABLE 16.2. Key Contractual Elements of an Infrastructural SLA

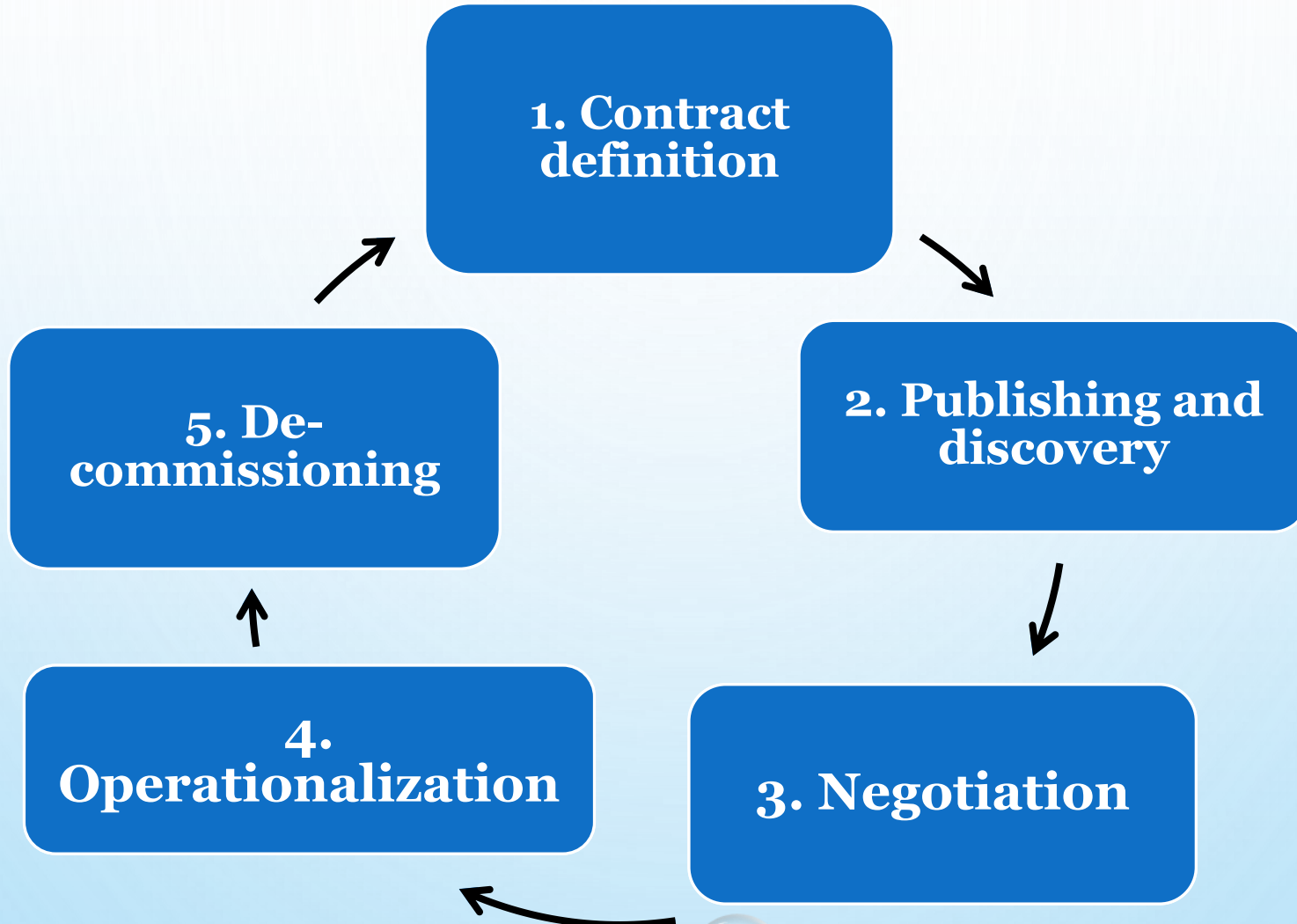
| | |
|--|---|
| <i>Hardware availability</i> | • 99% uptime in a calendar month |
| <i>Power availability</i> | • 99.99% of the time in a calendar month |
| <i>Data center network availability</i> | • 99.99% of the time in a calendar month |
| <i>Backbone network availability</i> | • 99.999% of the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

TYPES OF SLA

TABLE 16.3. Key contractual components of an application SLA

| | |
|---------------------------------------|--|
| <i>Service-level parameter metric</i> | <ul style="list-style-type: none"> • 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) |
| <i>Function</i> | <ul style="list-style-type: none"> • 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 |
| <i>Measurement directive</i> | <ul style="list-style-type: none"> • DB latency available via http://mgmtserver/em/latency • WS latency available via http://mgmtserver/ws/instanceno/latency |
| <i>Service-level objective</i> | <ul style="list-style-type: none"> • Service assurance |
| <i>Penalty</i> | <ul style="list-style-type: none"> • web site latency < 1 sec when concurrent connection < 1000 • 1000 USD for every minute while the SLO was breached |

LIFE CYCLE OF SLA



LIFE CYCLE OF SLA (CONTD..)

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.

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 (CONTD..)

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.

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 (CONTD..)

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

(CONTD..)

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, (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.
 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.

SLA MANAGEMENT IN CLOUD

(CONTD..)

● On-Boarding of Application

On-boarding activity consists of the following steps:

- a. Packing of the application for deploying on physical or virtual environments.
- b. The packaged application is executed directly on the physical servers to capture and analyze the application performance characteristics.
- c. The application is executed on a virtualized platform and the application performance characteristics are noted again.
- d. Based on the measured performance characteristics, different possible SLAs are identified.
- e. Once the customer agrees to the set of SLOs and the cost, the MSP starts creating different policies required by the data center for automated management of the application.
- f. Types of Policies: (1) business, (2) operational, and (3) provisioning

SLA MANAGEMENT IN CLOUD

(CONTD..)

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

(CONTD..)

Production:

- The application is made accessible to its end users under the agreed SLA.
- Additionally, customer may request the MSP for inclusion of new terms and conditions in the SLA.

Termination:

- When the customer wishes to withdraw the hosted application and does not wish to continue 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.



NEXT CLASS....