4.3 Traditional Approaches to SLO Management

SLO

- Service Level Objective
- An SLO (**service level objective**) is an agreement within an SLA about a specific metric like uptime or response time.
- So, if the SLA is the formal agreement between you and your customer, SLOs are the individual promises you're making to that customer.

What is an SLA?

- Users and providers need to clarify certain things to each other through a Service Level Agreement(SLA).
- An SLA is a contract between the user and the cloud service provider.
- It defines the terms of responsibility of the cloud service provider and the service credits if the provider is unable to meet the terms.

What is an SLA?

- It is critical document as data security, privacy, application availability are usually beyond the user's control.
- Cloud service delivers great value in terms of economy, but that should not diminish the value of an SLA.

SLA

- A service-level agreement (SLA) defines the level of service you expect from a vendor, laying out the metrics by which service is measured, as well as remedies or penalties should agreed-on service levels not be achieved.
- It is a critical component of any technology vendor contract.

SLA

- Gartner definition:
- A service-level agreement (SLA) sets the expectations between the service provider and the customer and describes the products or services to be delivered, the single point of contact for end-user problems, and the metrics by which the effectiveness of the process is monitored and approved.

SLA

- Four risks according to Gartner
  - Sourcing contracts are not mature for all markets.
  - Clauses in the contract document are usually written to favor the provider.
  - The clauses are not transparent.
  - Contract documents do not have clear service commitments. Their responsibility is limited to their equipment and software. Does not contain quality of service and its implications on fee.

SLA

- Despite of SLAs and outage related penalties, outages still occurs like:
  - Hard disks with rotating platters and heads crash
  - Viruses and malware that sometimes circumvent the IDS, IPS, UTM and firewalls.
  - Operators may inadvertently cause hardware and software glitches.

- Example:
- https://azure.microsoft.com/en-in/support/legal/sla/summary/

Need of SLA

- SLAs are an integral part of an IT vendor contract.
- An SLA pulls together information on all of the contracted services and their agreed-upon expected reliability into a single document.
- They clearly state metrics, responsibilities and expectations so that, in the event of issues with the service, neither party can plead ignorance.
- It ensures both sides have the same understanding of requirements.

Need of SLA

- Any significant contract without an associated SLA (reviewed by legal counsel) is open to deliberate or inadvertent misinterpretation.
- The SLA protects both parties in the agreement.
- Ideally, SLAs should be aligned to the technology or business objectives of the engagement.
- Misalignment can have a negative impact on deal pricing, QoS delivery, and customer experience.

Who provides the SLA?

- Most service providers have standard SLAs — sometimes several, reflecting various levels of service at different prices — that can be a good starting point for negotiation.
- These should be reviewed and modified by the customer and legal counsel, however, since they are usually slanted in favor of the supplier.

Who provides the SLA?

- When sending out an RFP, the customer should include expected service levels as part of the request; this will affect supplier offerings and pricing and may even influence the supplier's decision to respond.
- For example, if you demand 99.999 percent availability for a system, and the supplier is unable to accommodate this requirement with your specified design, it may propose a different, more robust solution.

What's in an SLA?

- The SLA should include not only a description of the services to be provided and their expected service levels, but also metrics by which the services are measured, the duties and responsibilities of each party, the remedies or penalties for breach, and a protocol for adding and removing metrics.

What's in an SLA?

- Metrics should be designed so bad behavior by either party is not rewarded.
- For example, if a service level is breached because the client did not provide information in a timely manner, the supplier should not be penalized.

Types of SLA

- Off the self SLAs
  - Available on their website.
  - They offer credit towards the monthly wanting to host critical services on the cloud.
  - Non-negotiable
- Negotiable SLAs
  - Customized as per customer needs.

What are key components of an SLA?

- SLA document contains
  - Service Level Objectives
    - Defines the characteristics of a service in specific and quantifiable terms.
  - Business Level Objectives
    - The basis for SLAs and SLOs.
    - Defines why the customer needs to use cloud computing.

What are key components of an SLA?

- The SLA should include components in two areas: services and management.
  - Service elements include specification of services provided,
  - conditions of service availability,
  - standards such as time window for each level of service,
  - responsibilities of each party, escalation procedures, and cost/service tradeoffs.

What are key components of an SLA?

- Examples:
  - The application must not have more than 15 pending requests at any instant.
  - Response for a read request should initiate within 3 seconds.
  - Data must be stored within the <XYZ> data centers.

What are key components of an SLA?

- Management elements should include
  - definitions of measurement standards and methods,
  - reporting processes, contents and frequency,
  - a dispute resolution process,
  - an indemnification clause protecting the customer from third-party litigation resulting from service level breaches (this should already be covered in the contract, however), and
  - a mechanism for updating the agreement as required.

What are key components of an SLA?

- This last item is critical; service requirements and vendor capabilities change, so there must be a way to make sure the SLA is kept up-to-date.

An SLA must contains:

- List of services the provider offered to you along with a definition of each service.
- Easy to understand metrics to evaluate if the provider is delivering the service at the promised levels.
- Mechanism to monitor the service.

An SLA must contains:

- The SLA must have easy-to-understand metrics to measure performance and availability.
  - Network and storage throughput
  - Application response speed
  - Maximum number of outages per month
  - Load based and dynamic elasticity are important, then the ability to add or remove resources in real-time must be an SLA requirement.

An SLA must contains:

- Customer responsibilities such as using licensed and tested applications on IaaS Virtual Machines, storing legitimate and virus-free data, not attempting to break-in to other tenants VM or accounts.
- Remedies or credits to be given if the term of the SLA are not met.
- Expected changes in the SLA over time.

An SLA must contains:

- Customer must demand that they get the following rights from the cloud provider:
  - Assurance of service quality
  - Transparent information on financial state of the cloud provider compliance to regulatory requirement.

SLA Aspects and Requirement

- Some considerations that must be specified within an SLA.
- Helps to make a compact SLA

Service Availability

- The SLA document must have information about the service uptime.
  - Promised uptime
  - Different based on type of application like mission critical services need 99.99% uptime
  - Specify what if uptime is lower than the one in the SLA?
  - Specify How provider inform you about the uptime. i.e Monthly, Yearly

Service Availability

- Specify minimum outage duration to qualify as downtime. i.e 5 min, 15 min
- Specify how downtime calculated in case longer duration.
  - Averaging downtime
- Note that downtime should be for user service or data and not for a component such as server, storage, connectivity, database or application.

- What if storage and data is down and server is up?

Data Locations

- The SLA must specify data locations.
- Decision taken based on
  - Sensitive data
  - Country's law
- You should have the right to visit and audit the attribute of the data center such as physical and network security, Disaster Recovery strategies, maintenance processes, etc.

Availability Zones (AZs)

- Data replicated to different availability zones.
- In some cases, the SLA penalty and outage are applicable only if all AZs are down.
- If data are not replicated, even though all AZs are down, not eligible for penalty.

Downtime Credits

- The provider may put a cap on the percentage of a customer's bill that can be reduced for downtime credits.
- The credit, if capped, are usually meager and less than the hard and soft losses such as lost sales opportunity, goodwill, brand image, morale, or productivity.

Credit Initiation

- Note who has the burden of initiating a credit.
- Initiate credit within specific time of receiving bill.
- Note the credit processing time.
- Determine when credit is reflected, next bill or after 6 Month.

Mean Time To Repair (MTTR)

- Insist that your provider give you an MTTR in the SLA.
- If time taken is more than MTTR, the provider must issue credit for extra time taken.

Data Protection

- The SLA should specify details of backups.
  - Frequency of backup
  - Storing tapes offsite
  - Data replicated to remote site

Data Encryption

- Encryption detail when
  - Data at rest
  - In motion
- Specify details of encryption procedures and access policies

Regulatory Requirements

- Based on enterprise regulatory need, specify in the SLA, such as
  - Data retention
  - Encryption
  - Data privacy
  - Authentication and authorization policies
- Must be transparent and help you during your compliance audits.

Certifications

- Important for compliance.
- The SLA must specify that the provider has and will maintain certain certifications.
- Such as Payment Card Industry Data Security Standards (PCIDSS), Health Insurance Portability and Accountability (HIPAA)

Advance Notification

- The SLA must include:
  - The notification in advance for any scheduled maintenance and downtime.
- After an issue is discovered, providers must share information about security breaches, regardless of whether the breach impacted your data or service.
- Your data is subject to the same risk, you must be aware of what is happening so you can implement measures to secure your data.

Scheduled Maintenance Periods

- The SLA must specify if the service will be available during scheduled maintenance period.

Closer Notice Period

- The SLA must specify closure notice period.
- So that you can migrate your data, service etc.
- Be aware of the local laws where the data center and cloud provider located.
- Cloud provider should give you enough notice to save or migrate your data to enterprise or other cloud in case
  - Law Enforcement Agencies seize the property
  - Cloud provider goes bankrupt

Hidden Cost

- Read the SLA for hidden cost.
- Investigation cost in case of issue reported by you.
- Specify how charges will be applied.
- Upper limit of bill in such case.

Floating Terms

- Policies and terms may be published on website.
- Sometimes cloud provider may nee the flexibility to change some SLA terms.
- Specify how and what can be change and not change.
- Easy exit clause if new terms are not acceptable.

SLA-All to gather

- An SLA should be beneficial for both the customer and the cloud provider.
- It should be the best solution for the business.
- Always prepare an SLA that is balanced and a win-win for all party.
- When reviewing an SLA, gather all the concerned involved from security, IT, business divisions, legal accounting, etc.

4.4 Traditional Approaches to SLO Management

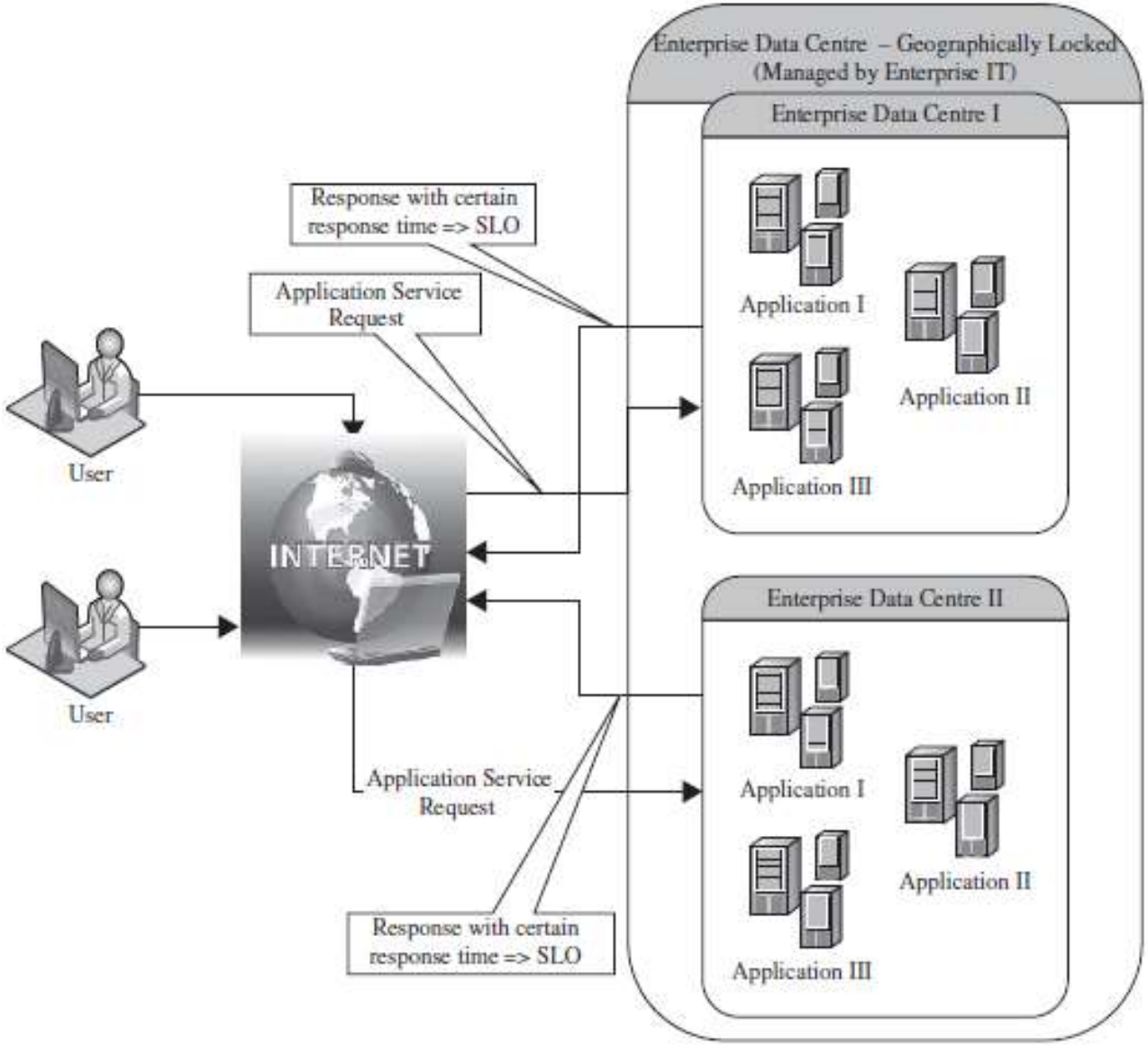
Background

- The main reason for provisioning server resource – peak load
- Process is tedious.
- Web application hosted on server – dedicated server room – provides e-services to various clients.

Background

- Typical Service Level Objectives for these applications were response time and throughput of the application end-user request.
- Capacity planning – determining servers and their capacity during peak load.

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



Background

- The number of web applications and their complexity have grown.
- Therefore, the complexity of managing the data centers also increased.
- Accordingly, enterprises realized that it was economical to outsource the application hosting activity to third-party infrastructure providers because:

Background

- 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.

Background

- 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).

Background

- These SLAs are known as the infrastructure SLAs.
- The infrastructure service providers are known as Application Service Providers (ASPs).
- Consequently, a set of tools for monitoring and measurement of availability of the infrastructure were required and developed.

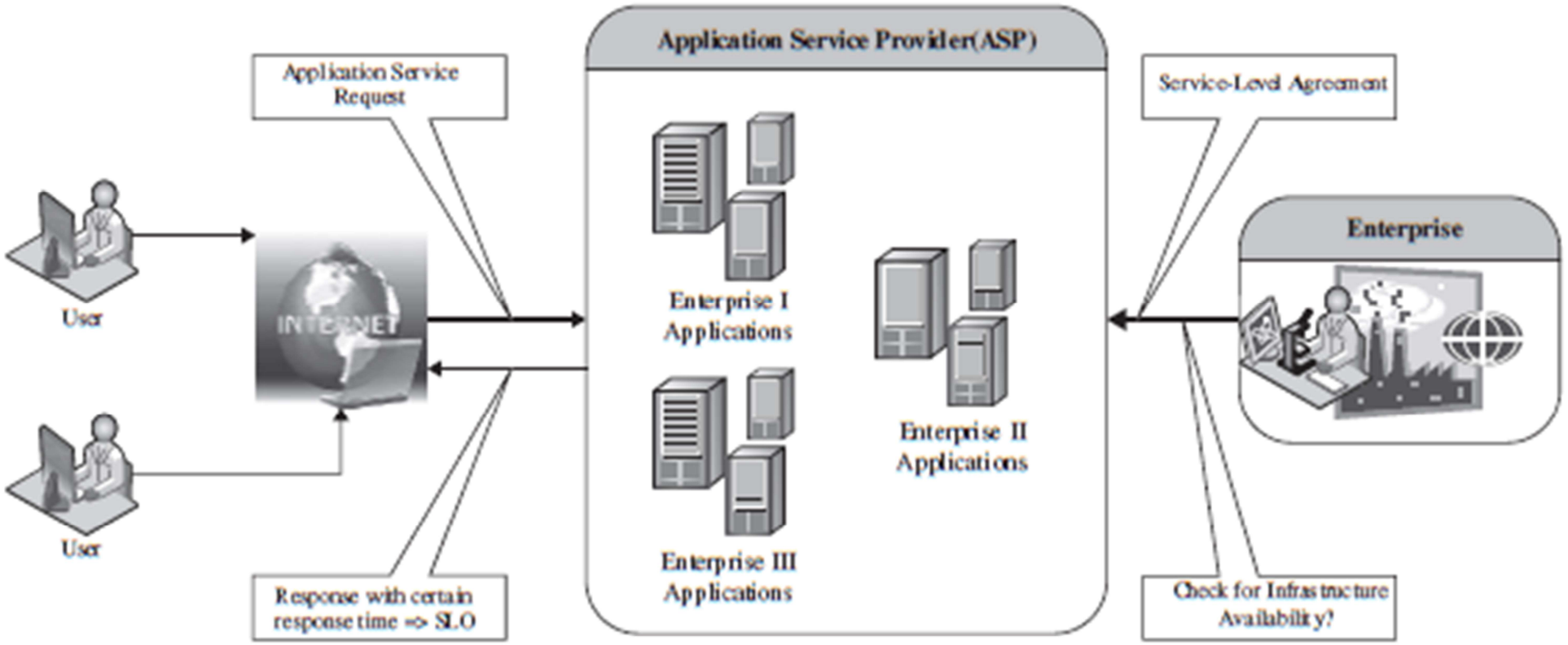
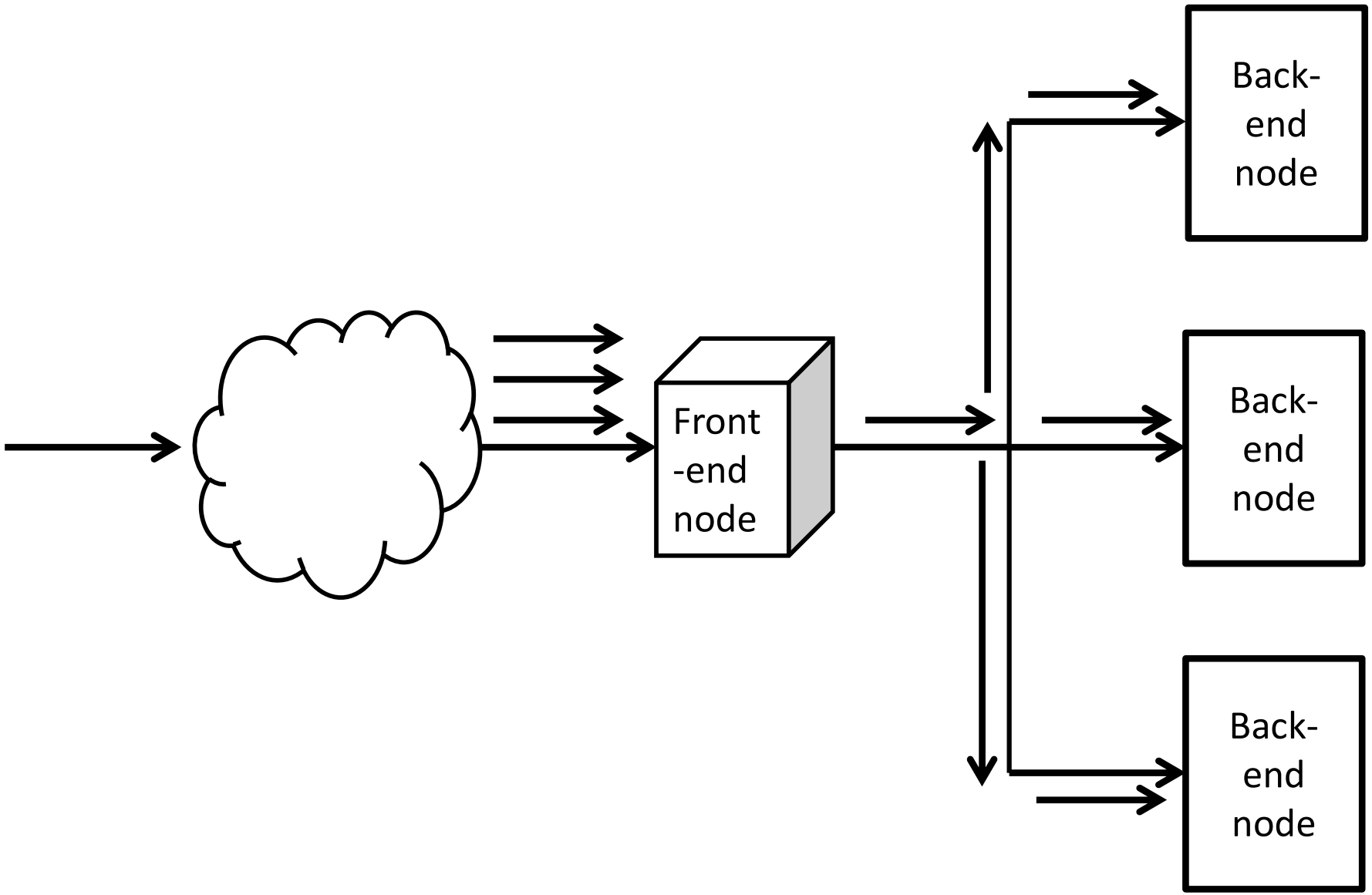


FIGURE 16.2. Dedicated hosting of applications in third party data centers.

- 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.
- This led to further cost advantage for both the ASPs and enterprises.

Background

- First attempt to SLO management
  - To provide guaranteed QoS for hosted web applications
    - Load balancing techniques
    - Admission control mechanisms


Front
-end
node
Back-
end
node
Back-
end
node
Back-
end
node

Load Balancing

- Objective
  - To distribute the incoming requests onto a set of physical machines, each hosting a replica of an application, so that load on the machines is equally distributed.
  - Load balancing algorithm runs on physical machine that interfaces with the clients also called as front-end node.
  - Machine serving incoming requests are known as the back-end nodes.

4.5 Life Cycle of SLA

Life Cycle of SLA

- Each SLA goes through 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 five phases:


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

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.

Publishing and Discovery

- Service provider advertises these base service offerings through standard publication media.
- 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.

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.

Negotiation

- 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.

Negotiation

- 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.

Operationalization

- SLA operation consists of
  - SLA monitoring,
  - SLA accounting, and
  - SLA enforcement

Operationalization

- **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.

Operationalization

- **SLA accounting** involves capturing and archiving the SLA adherence for compliance.
- As part of accounting, the application's actual performance and the performance guaranteed as a part of SLA is reported.

Operationalization

- Apart from the frequency and the duration of the SLA breach, it should also provide the penalties paid for each SLA violation.
- **SLA enforcement** involves taking appropriate action when the runtime monitoring detects a SLA violation.
- Such actions could be notifying the concerned parties, charging the penalties besides other things.

Operationalization

- The different policies can be expressed using a subset of the Common Information Model (CIM) .
- The CIM model is an open standard that allows expressing managed elements of data center via relationships and common objects.

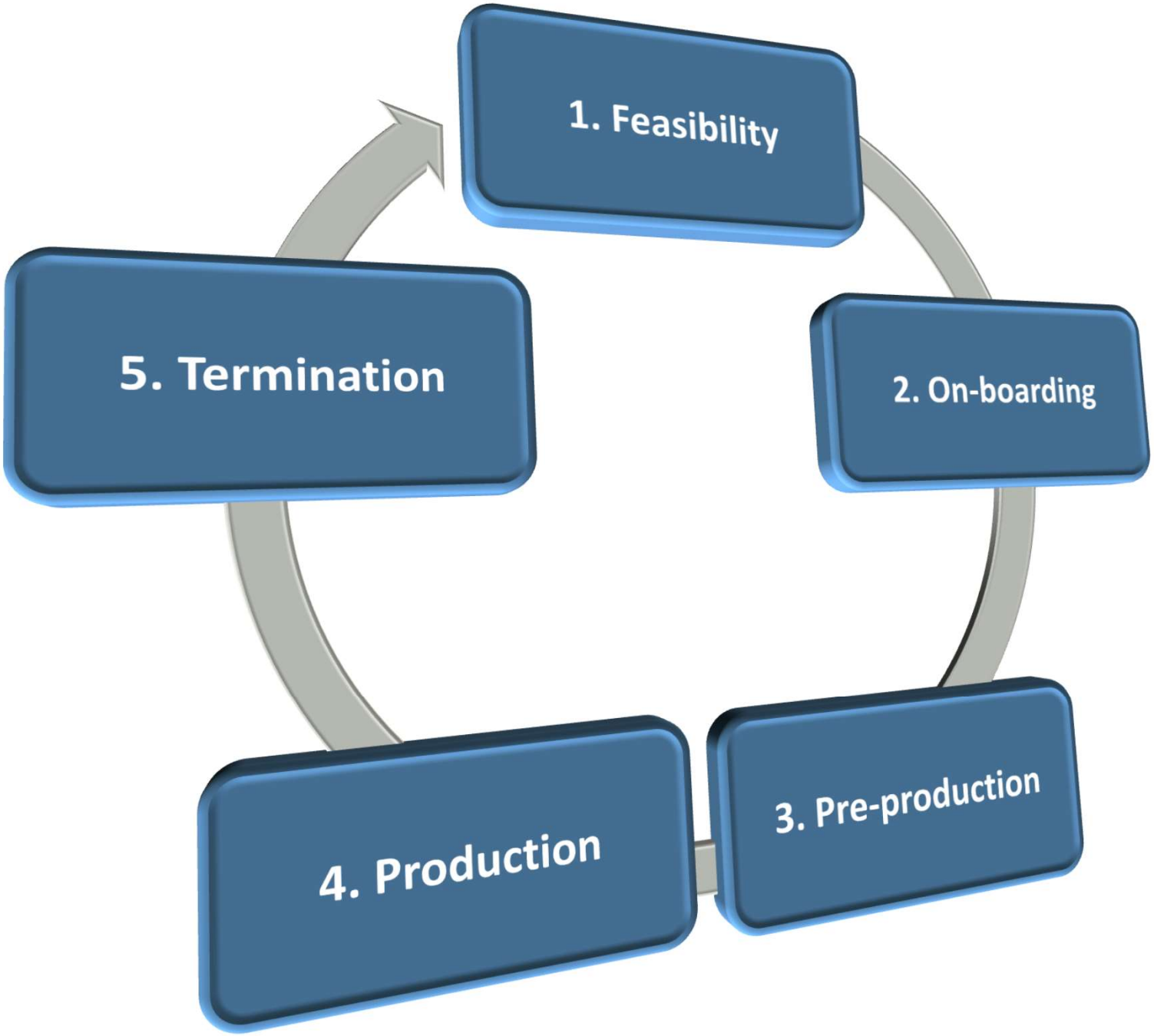
De-commissioning

- 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.

4.6 SLA Management in Cloud

Life Cycle of SLA

- SLA management of applications hosted on cloud platforms involves five phases.


1. Feasibility
2. On-boarding
3. Pre-production
4. Production
5. Termination

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

Feasibility Analysis

Technical feasibility

- Ability of an application to scale out.
- Compatibility of the application with the cloud platform being used within the MSP's data center.
- The need and availability of a specific hardware and software required for hosting and running of the application.
- Preliminary information about the application performance and whether they can be met by the MSP.

Feasibility Analysis

Infrastructure feasibility

- Availability of infrastructural resources in sufficient quantity so that the projected demands of the application can be met.

Feasibility Analysis

Financial feasibility

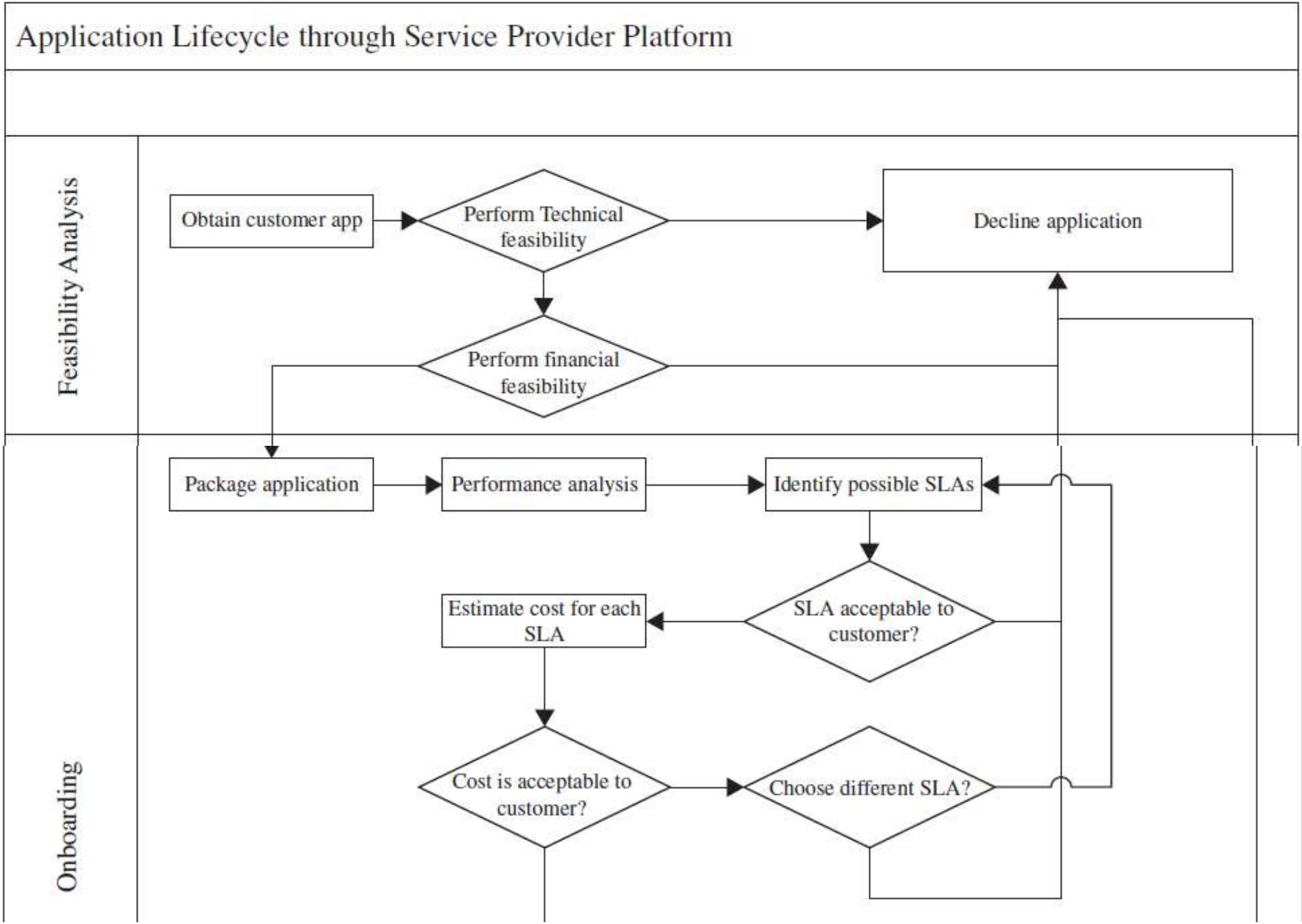
- Determining the approximate cost to be incurred by the MSP and the price the MSP charges the customer so that the hosting activity is profitable to both of them.

Feasibility Analysis

- A feasibility report consists of the results of the above three feasibility studies.
- The report forms the basis for further communication with the customer.
- Once the provider and customer agree upon the findings of the report, the outsourcing of the application hosting activity proceeds to the next phase, called “onboarding” of application.

Feasibility Analysis

- Only the basic feasibility of hosting an application has been carried in this phase.
- However, the detailed runtime characteristics of the application are studied as part of the on-boarding activity.


Application Lifecycle through Service Provider Platform
Feasibility Analysis
Obtain customer app
Perform Technical feasibility
Decline application
Perform financial feasibility
Onboarding
Package application
Performance analysis
Identify possible SLAs
SLA acceptable to customer?
Estimate cost for each SLA
Cost is acceptable to customer?
Choose different SLA?

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.

On-Boarding of Application

- 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 onboarding activity is completed.

On-Boarding of Application

- The on-boarding activity consists of:
  - Packing of the application for deploying on physical or virtual environments.
    - Application packaging is the process of creating deployable components on the hosting platform (could be physical or virtual).
    - Open Virtualization Format (OVF) standard is used for packaging the application for cloud platform

On-Boarding of Application

- Capture and analyze the application performance characteristics.
  - To validate customer's application.
  - it provides a baseline performance value for the application in non-virtual environment.
  - It helps to identify the nature of application—that is, whether it is CPU-intensive or I/O intensive or network-intensive and the potential performance bottlenecks.

On-Boarding of Application

- Important performance characteristics like the application's ability to scale (out and up) and performance bounds (minimum and maximum performance) are noted.

On-Boarding of Application

- Different possible SLAs are identified based on measured performance characteristics.
  - The resources required and the costs involved for each SLA are also computed.

On-Boarding of Application

- Once agreed on SLAs, MSP starts creating different policies required by the data center for automated management of the application.
  - This implies that the management system should automatically infer the amount of system resources that should be allocated/de-allocated to/from appropriate components of the application when the load on the system increases/decreases.
  - These policies are of three types: (1) business, (2) operational, and (3) provisioning.

On-Boarding of Application

- Business policies
  - help prioritize access to the resources in case of contentions.
  - Business policies are in the form of weights for different customers or group of customers.
- Operational policies
  - are the actions to be taken when different thresholds/conditions are reached.

On-Boarding of Application

- Also, the actions when thresholds/ conditions/triggers on service-level parameters are breached or about to be breached are defined.
- The corrective action could be different types of provisioning such as scale-up, scale-down, scale-out, scale-in, and so on, of a particular tier of an application.
- Additionally, notification and logging action (notify the enterprise application’s administrator, etc.) are also defined.

On-Boarding of Application

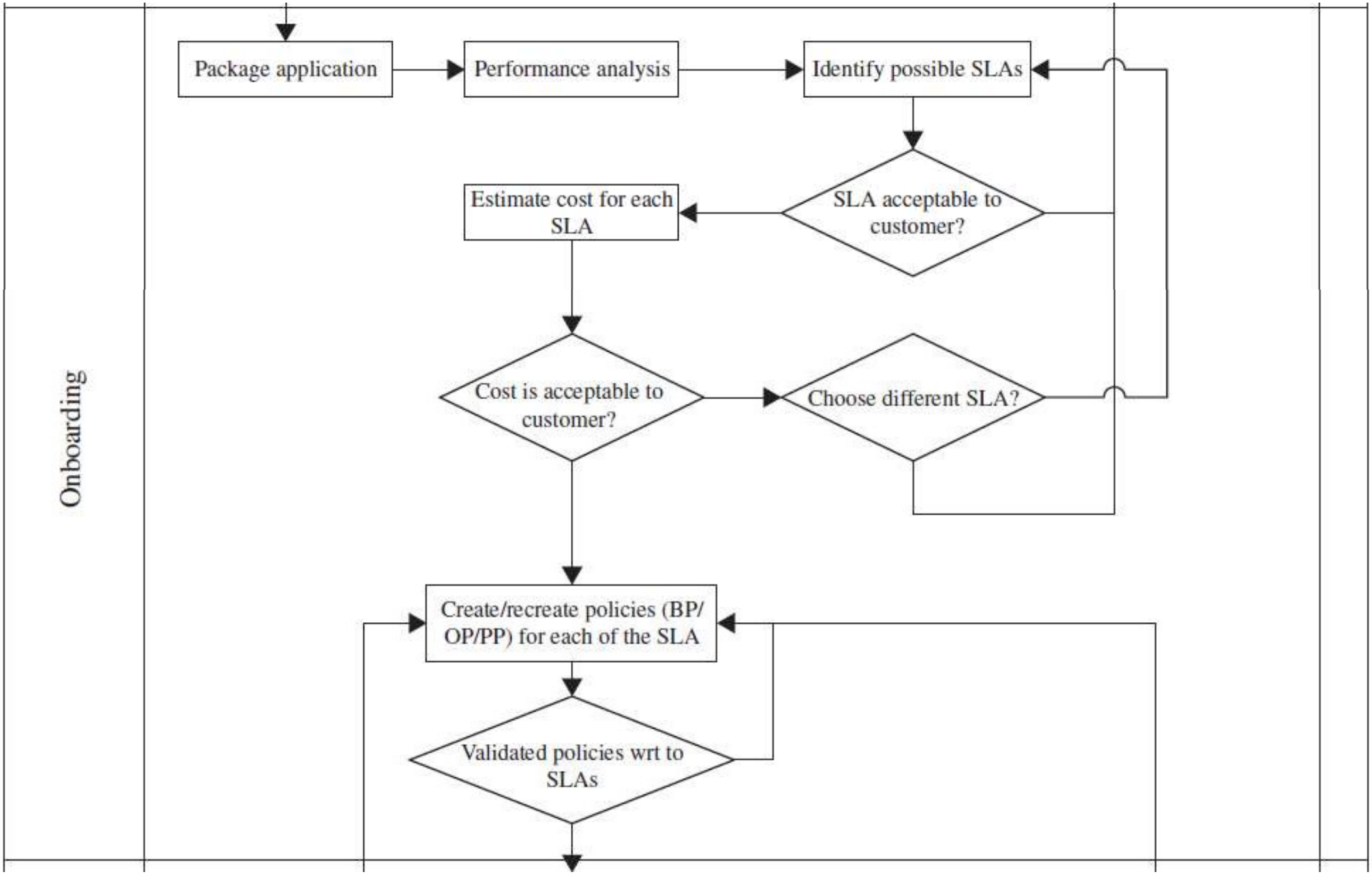
- Operational policies (OP) are represented in the following format:
- OP = collection of <Condition, Action>
- For example, one OP is
- OP = <average latency of web server . 0.8 sec, scale-out the web-server tier>

On-Boarding of Application

- Provisioning policies
  - help in defining a sequence of actions corresponding to external inputs or user requests.
  - Scale-out, scale-in, start, stop, suspend, resume are some of the examples of provisioning actions.
  - A provisioning policy (PP) is represented as
  - PP = collection of <Request, Action>
  - For example, a provisioning policy to start a web site consists of the following sequence:
  - start database server, start web-server instance 1, followed by start the web-server instance 2, and so on.

On-Boarding of Application

- On defining these policies, the packaged applications are deployed on the cloud platform and the application is tested to validate whether the policies are able to meet the SLA requirements.
- This step is iterative and is repeated until all the infrastructure conditions necessary to satisfy the application SLA are identified.
- Once the different infrastructure policies needed to guarantee the SLOs mentioned in the SLA are completely captured, the on-boarding activity is said to be completed.

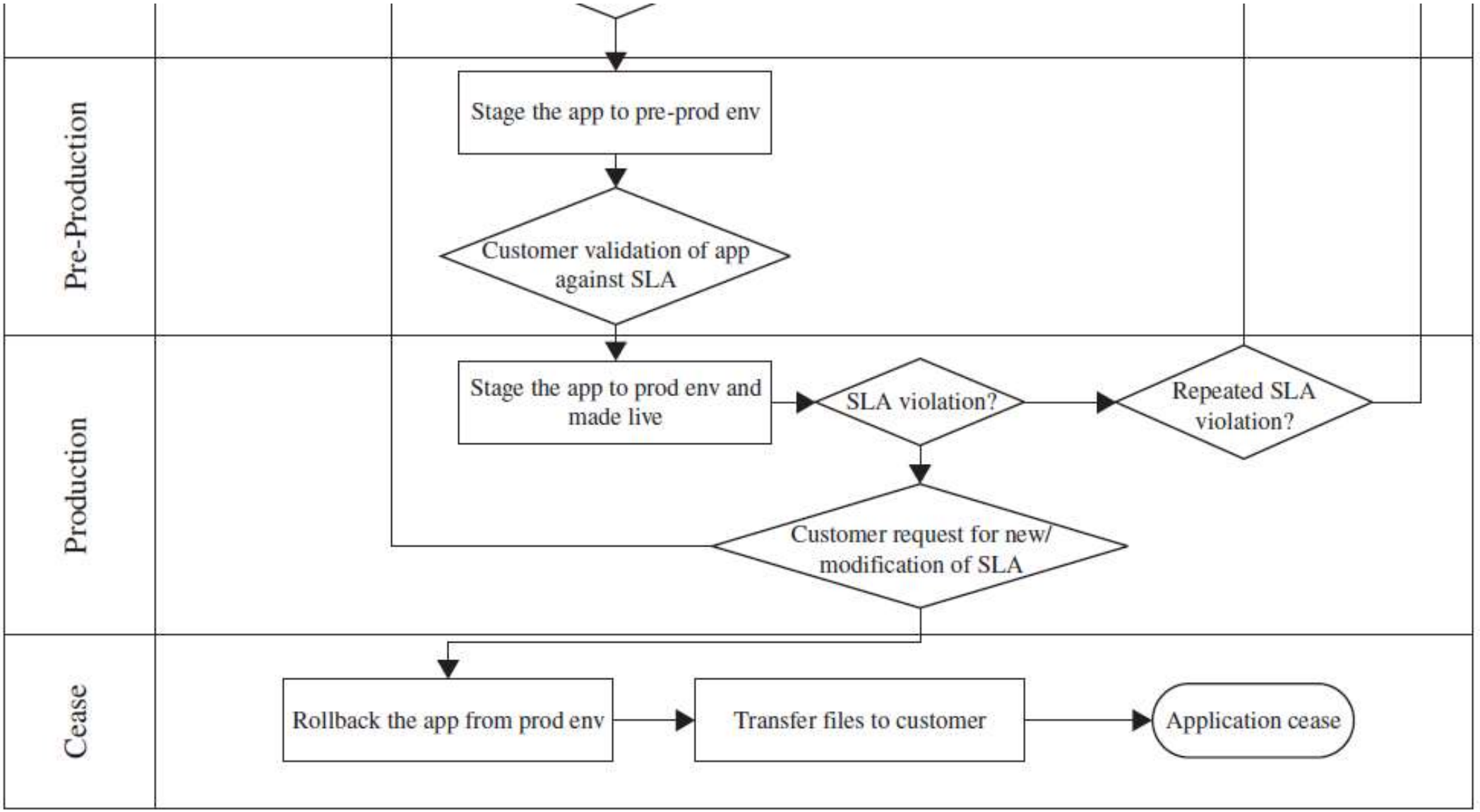

Onboarding
Package application
Performance analysis
Identify possible SLAs
SLA acceptable to customer?
Estimate cost for each SLA
Cost is acceptable to customer?
Choose different SLA?
Create/recreate policies (BP/ OP/PP) for each of the SLA
Validated policies wrt to SLAs

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.

Preproduction

- 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.


Pre-Production
Production
Cease
Stage the app to pre-prod env
Customer validation of app against SLA
Stage the app to prod env and made live
SLA violation?
Repeated SLA violation?
Customer request for new/ modification of SLA
Rollback the app from prod env
Transfer files to customer
Application cease

Production

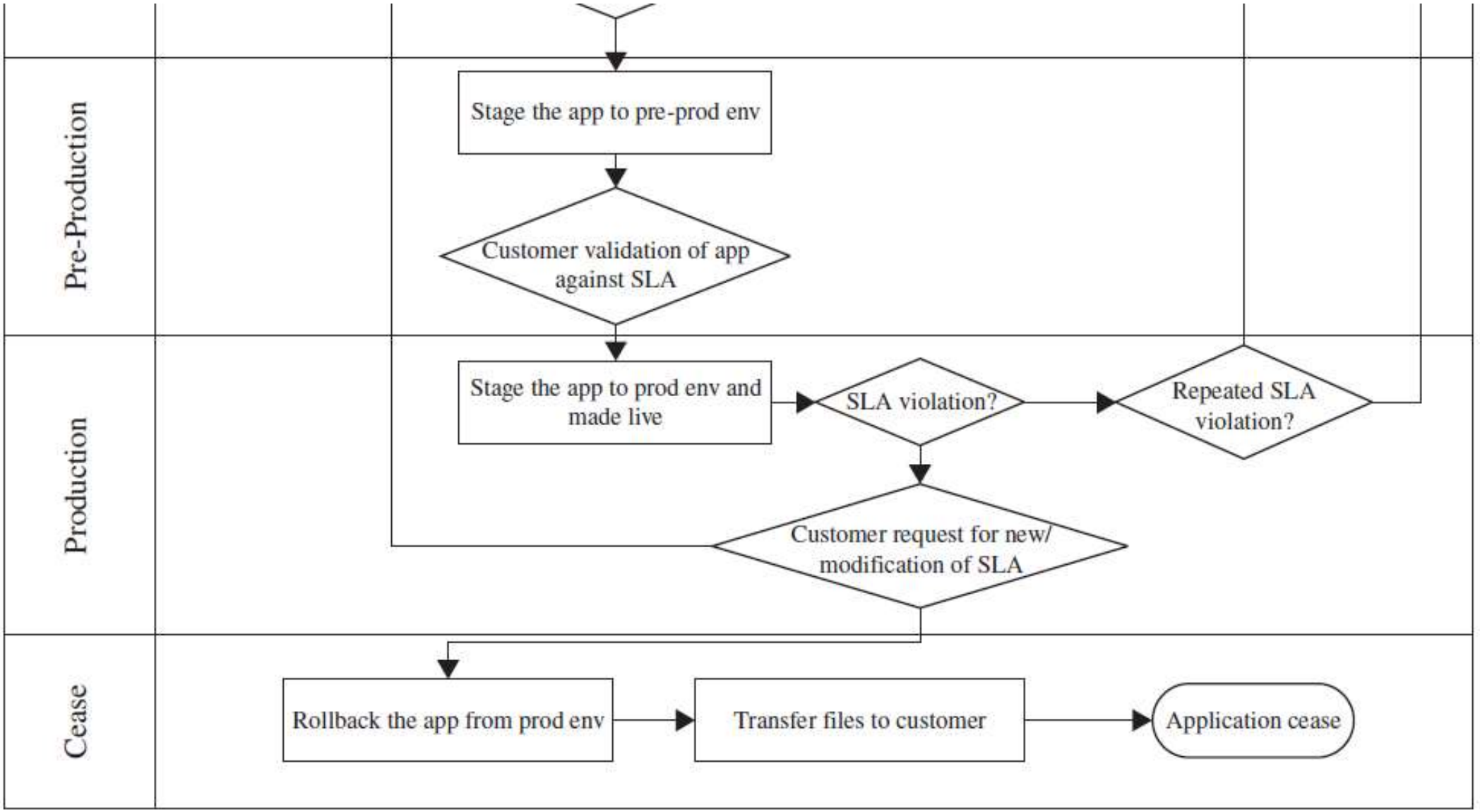
- In this phase, the application is made accessible to its end users under the agreed SLA.
- However, there could be situations when the managed application tends to behave differently in a production environment compared to the preproduction environment.
- This in turn may cause sustained breach of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

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.

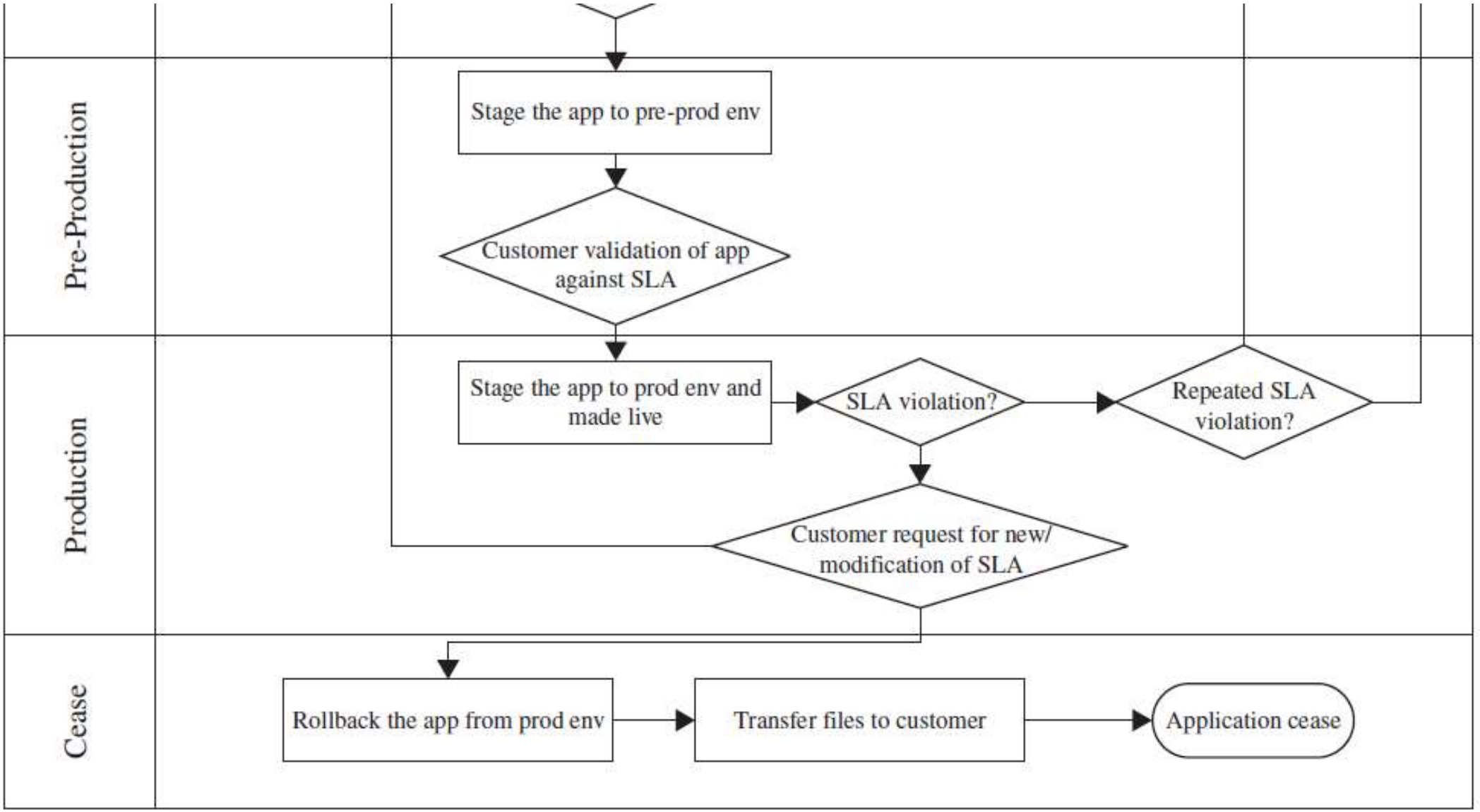
Production

- In the case of the former, on-boarding activity is repeated to analyze 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.


Pre-Production
Production
Cease
Stage the app to pre-prod env
Customer validation of app against SLA
Stage the app to prod env and made live
SLA violation?
Repeated SLA violation?
Customer request for new/ modification of SLA
Rollback the app from prod env
Transfer files to customer
Application cease

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.


Pre-Production
Production
Cease
Stage the app to pre-prod env
Customer validation of app against SLA
Stage the app to prod env and made live
SLA violation?
Repeated SLA violation?
Customer request for new/ modification of SLA
Rollback the app from prod env
Transfer files to customer
Application cease

NEXT: AUTOMATED POLICY BASED MANAGEMENT

4.7 Automated Policy based Management

Automated Policy based Management

- In on-boarding activities Operational and Provisioning policy 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.

Automated Policy based Management

- 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

- Consider a three tier Web application consisting of:
  - Web server
  - Application server
  - Database server

- How to decide resource requirement?

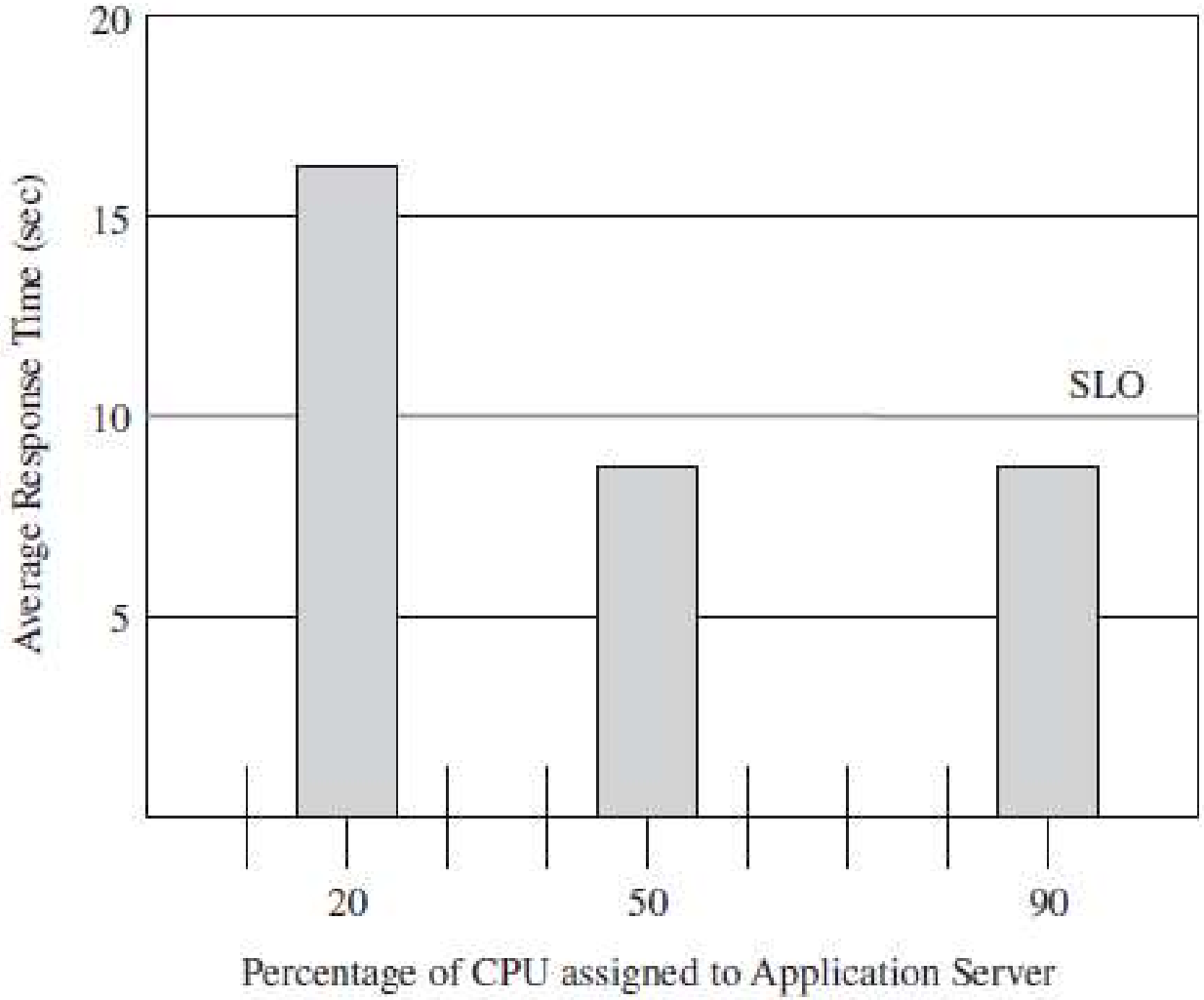


FIGURE 16.8. Performance of a multi-tier application for varied CPU allocation.

Automated Policy based Management

- To understand the system resource requirements for each of the tiers of an application at different workloads necessitates the deployment of the application on a test system.
- The test system is used to collect the low-level system metrics such as usage of memory and CPU at different workloads, as well as to observe the corresponding high-level service level objectives such as average response time.

Automated Policy based Management

- The metrics thus collected are used to derive the functional relationship between the SLOs and low-level system attributes.
- These functional relations are called policies.

Automated Policy based Management

- The triggering of operational and provisional policies results in a set of actions to be executed by the service provider platform.
- It is possible that some of these actions contend for the same resources.
- In such a case, execution of certain actions needs to be prioritized over the execution of others.

Automated Policy based Management

- The rules that govern this prioritization of request execution in case of resource contention are specified as a part of business policy.
- Some of the parameters often used to prioritize action and perform resource contention resolution are:

Automated Policy based Management

- 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.

Automated Policy based Management

- 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

- Priority ranking algorithms use these parameters to derive scores.
- These scores are used to rank each of the actions that contend for the same resources.
- Actions having high scores get higher priority and hence, receive access to the contended resources.

Automated Policy based Management

- Automatic operationalization of policies consists of following components:
  - Prioritization Engine
  - Provisioning Engine
  - Rules Engine
  - Monitoring System
  - Auditing
  - Accounting/Billing System

Automated Policy based Management

- 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.

Automated Policy based Management

- 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.

Automated Policy based Management

- 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.

Automated Policy based Management

- 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.

Automated Policy based Management

- 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.

Automated Policy based Management

- 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

- The policies and packaged application are deployed on the platform after completing the on-boarding activity.
- The customer is provided with options to start the application in any of the agreed SLAs.
- The application request is sent via the access layer to the system.

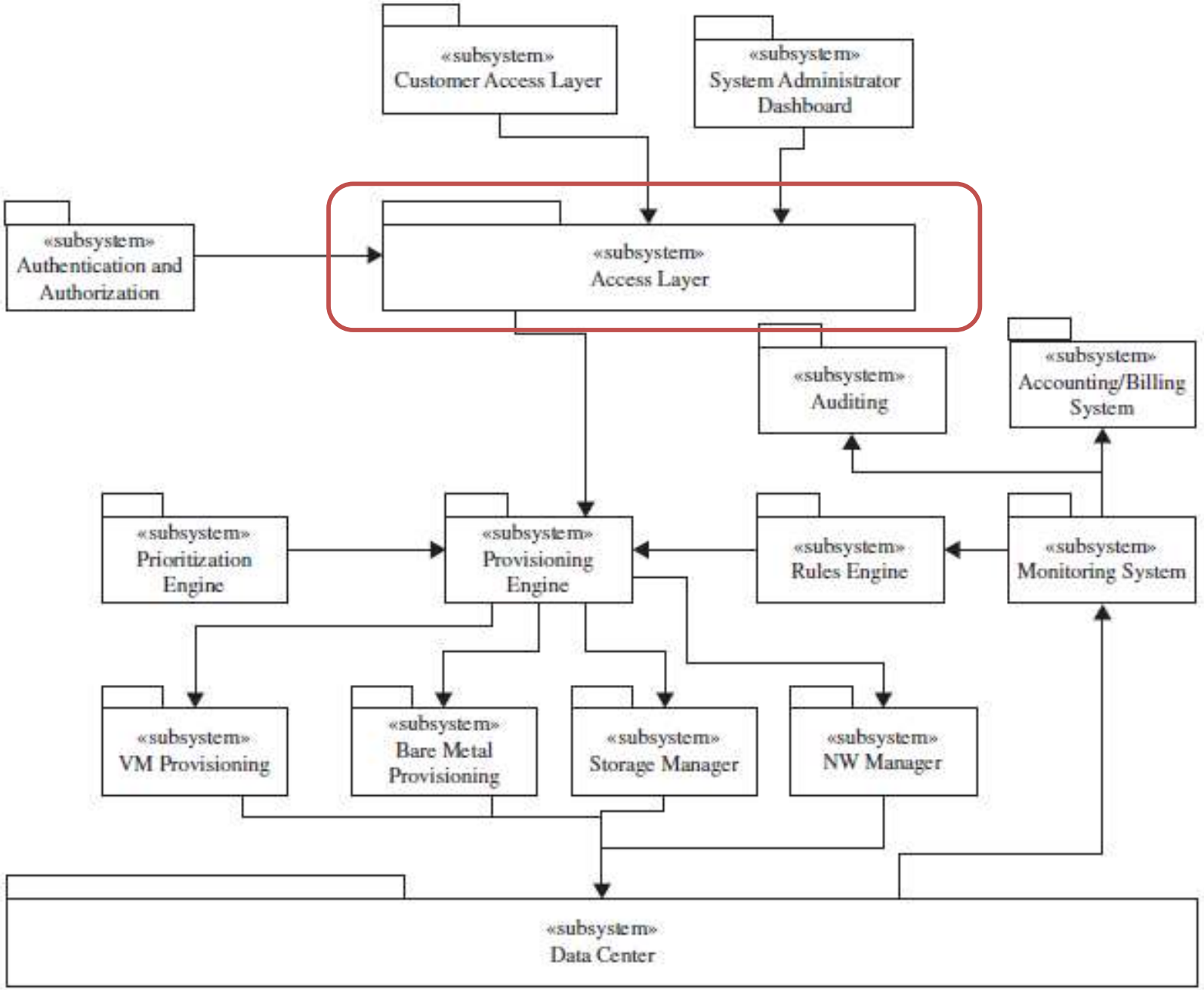


FIGURE 16.9. Component diagram of policy-based automated management system.

Automated Policy based Management

- Using the provisioning policy the provisioning engine determines how and in what sequence the different components/tiers of an application should be started and configured.

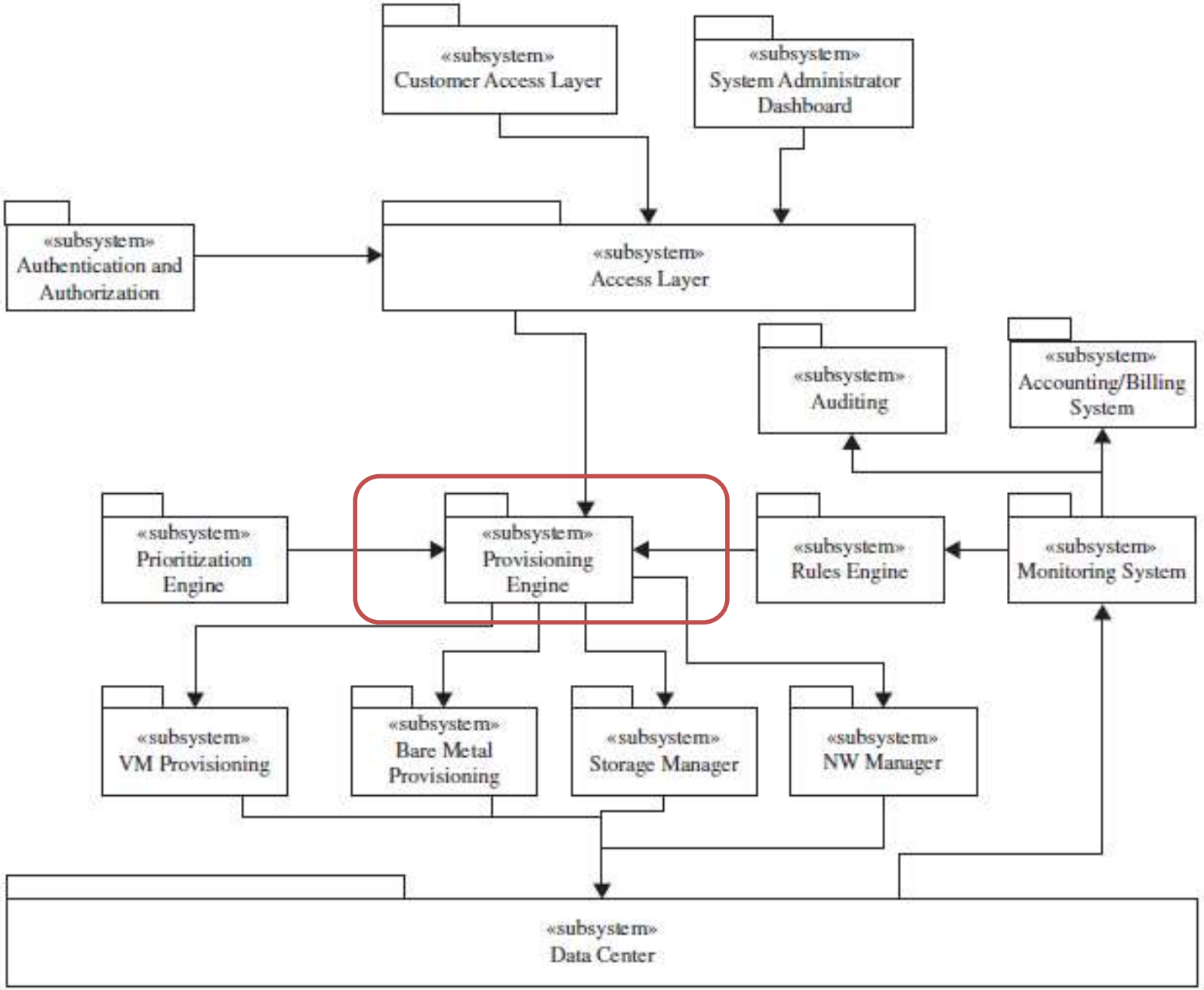


FIGURE 16.9. Component diagram of policy-based automated management system.

Automated Policy based Management

- If the start operation requires a resource that is also contended by a different application request, then provisioning engine interacts with the prioritization engine to determine the request that should have access to the contended resource in case of conflict.
- This conflict resolution is guided by the business policy defined in the prioritization engine.

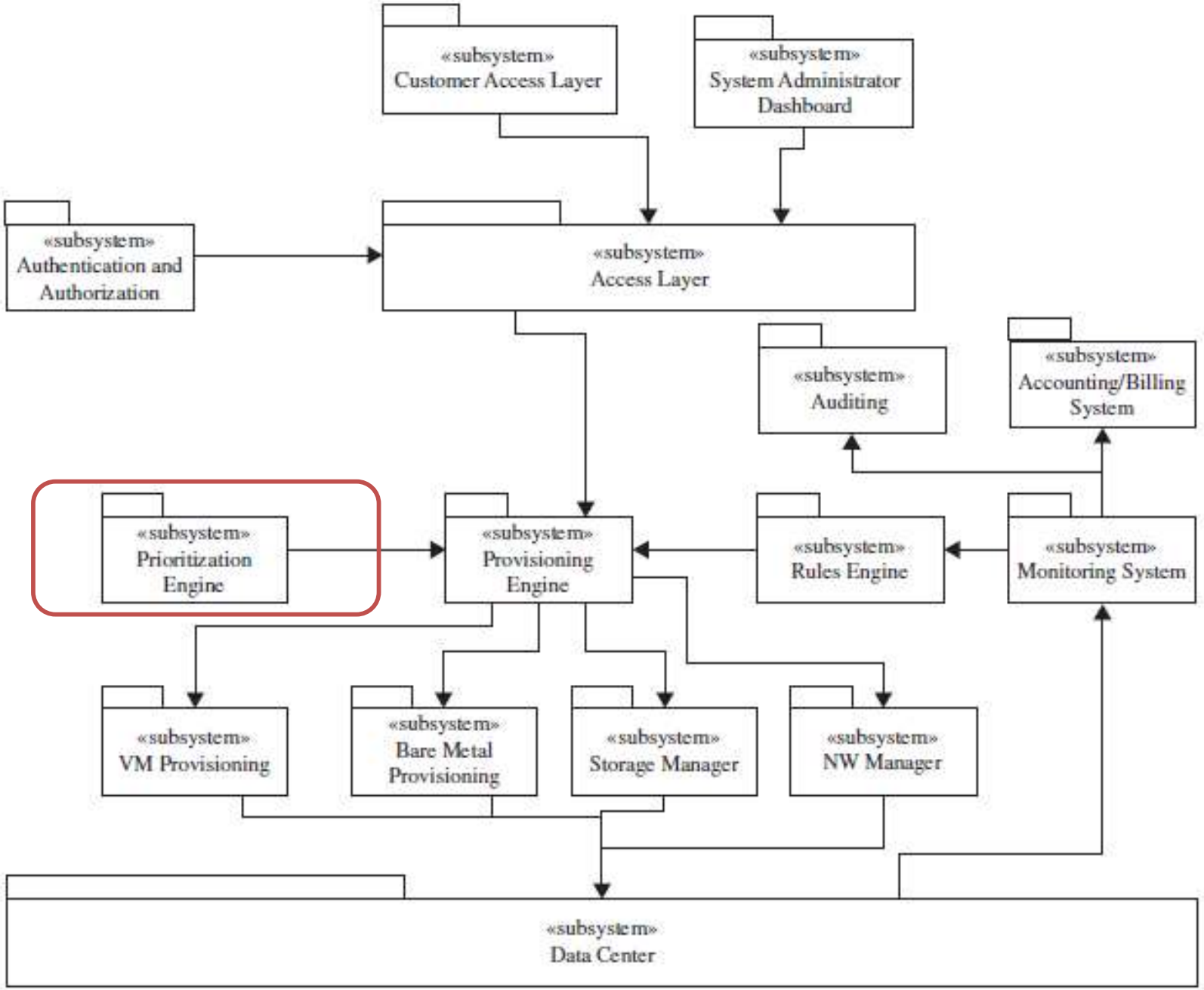


FIGURE 16.9. Component diagram of policy-based automated management system.

Automated Policy based Management

- Once an application begins execution, it is continuously monitored by the monitoring system.
- Monitoring involves collecting statistics about the key metrics and evaluating them against the rules defined in the operational policy for validating the SLA adherence.
- SLA violation triggers rules that initiate appropriate corrective action automatically.

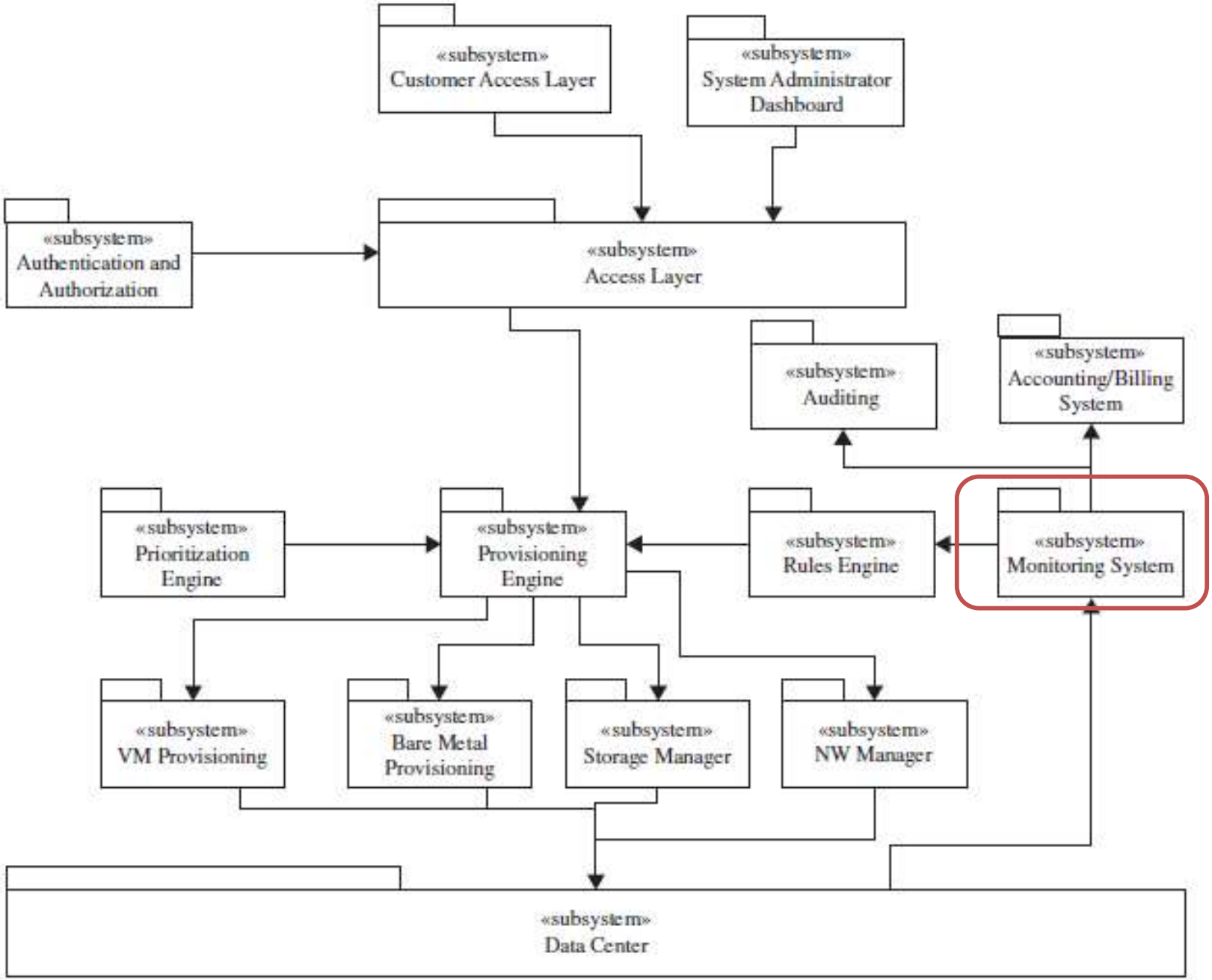


FIGURE 16.9. Component diagram of policy-based automated management system.

Automated Policy based Management

- For example, whenever the performance of the application degrades and chances of violating the agreed SLO limits are high, the rules that help scale out the bottleneck tier of the application is triggered.
  - This ensures that the performance does not degenerate to a level of violating the SLA.
- Periodically, the amount of resource utilized by the application is calculated.
  - On calculating the resource utilization, the cost is computed correspondingly and the bill is generated.
  - The bill along with the report on the performance of the application is sent to the customer.

Automated Policy based Management

- Alternatively, the monitoring system can interact with the rules engine through an optimization engine.
- The role of the optimization system is to decide the migration strategy that helps optimize certain objective functions for virtual machine migration.
- The objective could be to minimize the number of virtual machines migrated or minimize the number of physical machines affected by the migration process.

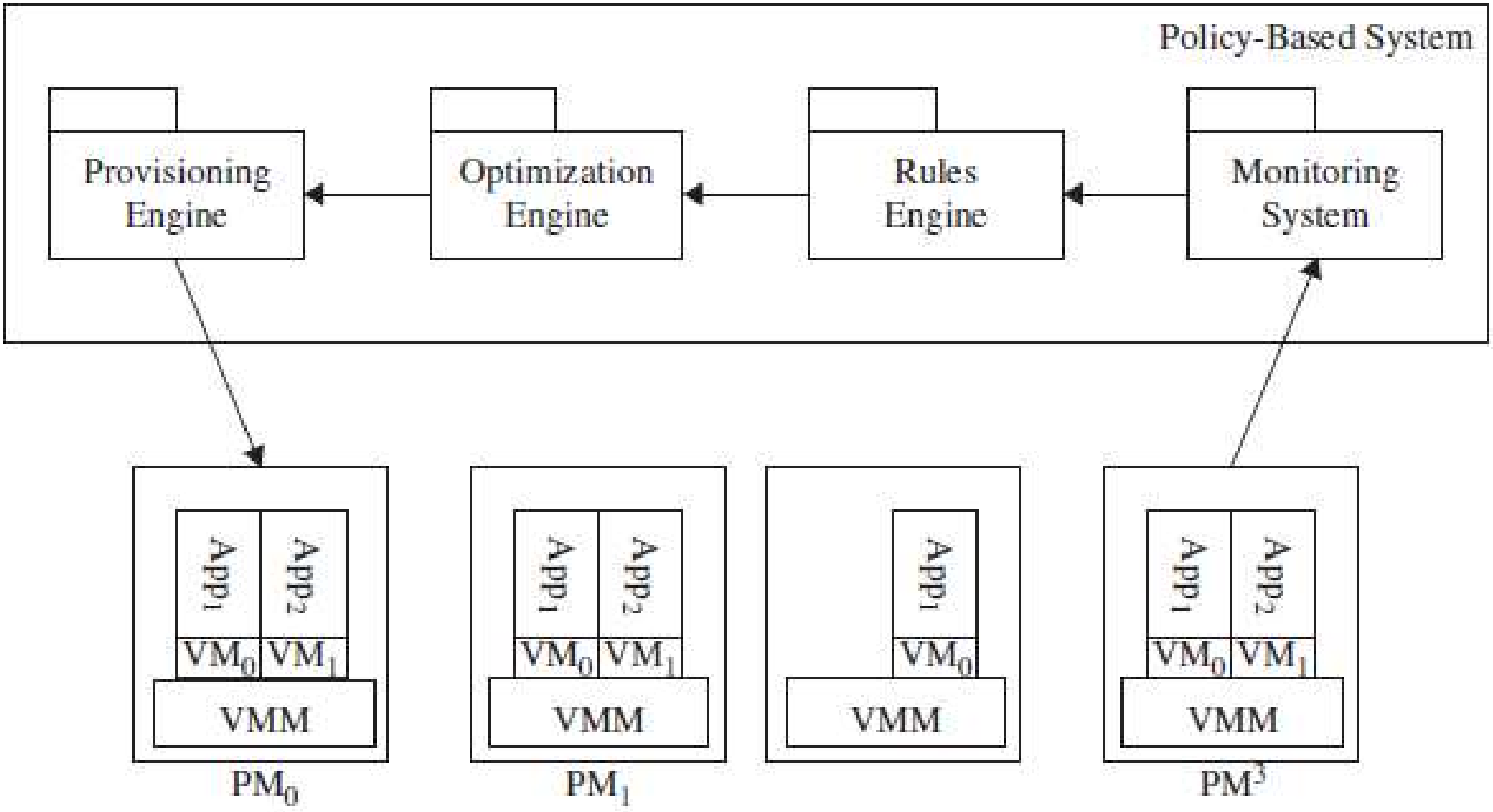
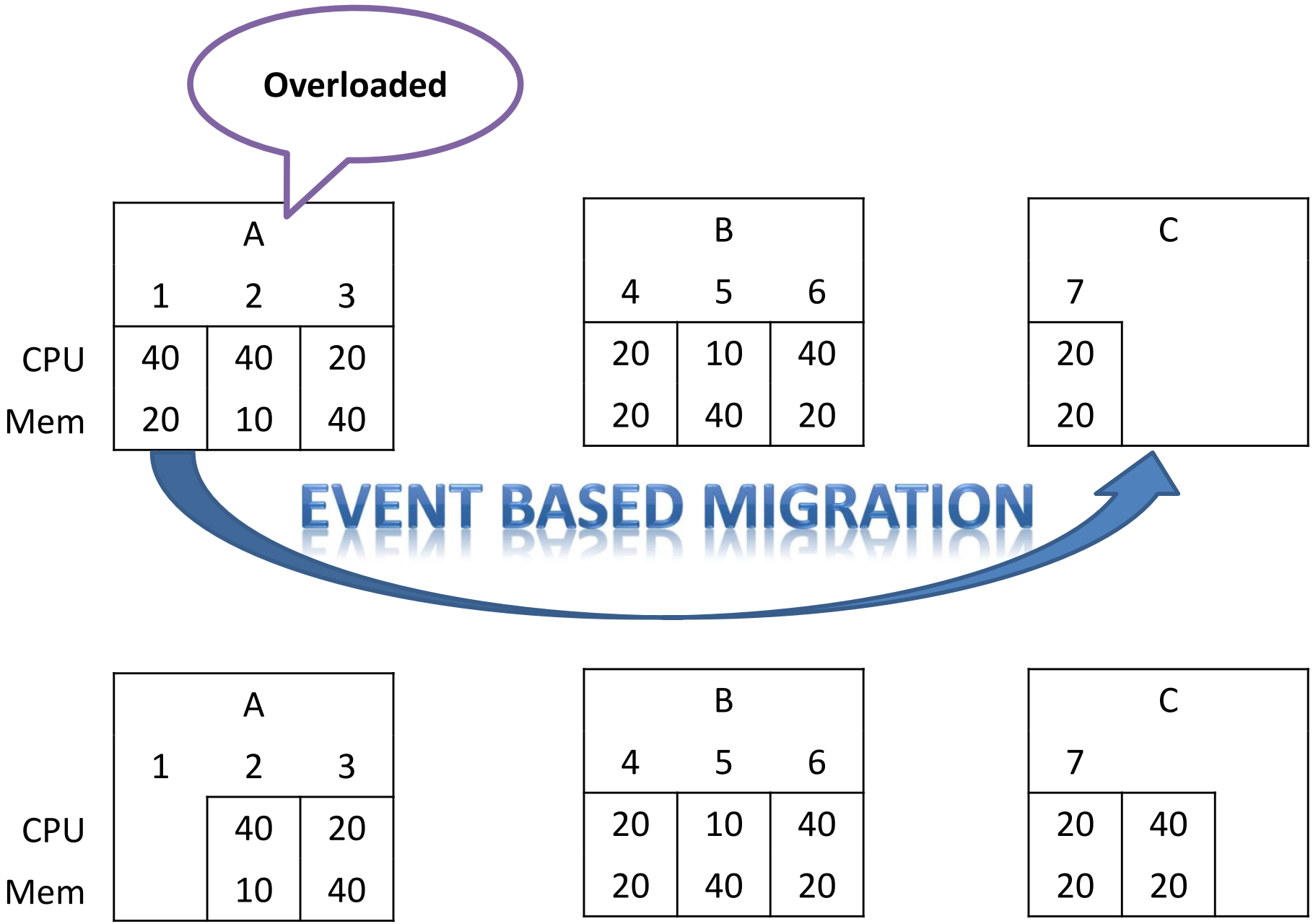


FIGURE 16.10. Importance of optimization in the policy-based management system.


Overloaded
A
1
2
3
CPU
40
40
20
Mem
20
10
40
B
4
5
6
20
10
40
20
40
20
C
7
20
20
EVENT BASED MIGRATION
A
1
2
3
CPU
40
20
Mem
10
40
B
4
5
6
20
10
40
20
40
20
C
7
20
40
20
20

Assume VM_4 require memory increased by 40
Result in overloaded -> Event based migration

	A		B			C		
	1	2	3	4	5	6	7	
CPU		40	20	20	10	40	20	40
Mem		10	40	40	40	20	20	20

	A		B			C		
	1	2	3	1	2	3	7	
CPU		40	20		10	40	20	40	20
Mem		10	40		40	20	20	20	40

Assume new VM_8 with CPU and memory requirement 70 need to allocate

	A		B		C		
	1	2	3	4	5	6	7		
CPU		40	20		10	40	20	40	20
Mem		10	40		40	20	20	20	40

Automated Policy based Management

- In this situation new machine D switched on for hosting because VM_8 cannot be hosted on any of the machine A, B and C.
- Now consider that, VM_1 and VM_4's QoS and SLA violations are well within the permissible limits.
- It require less number of PM to be switched on.

END