CHAPTER 6

CLOUD MANAGEMENT

6.1 INTRODUCTION

Companies and their IT vendors are focussed increasingly on virtualization-based cloud services and consolidation solutions, and the potential benefits they provide businesses. But this growing popularity sometimes obscures the fact that managing cloud virtualized infrastructures can present organizations significant challenges in areas related to implementation and service management. In particular, it is often difficult for businesses to determine how and for what purposes employees and groups are utilizing virtualized IT assets.

What is required is a cloud solution that aims to help overcome these problems by providing insights into the relationships between virtualized and physical IT assets – who is utilizing shared resources and what and how much they are using. Such information is critical in a number of ways. For organizations such as IT outsourcers, a well-suited solution will serve as an accurate measurement tool underlying billing processes and service level agreement (SLA) compliance.

Innovative cloud virtualization technologies from cloud vendors extend that concept far beyond simple partitioning on a single server to a systems virtualization platform. This platform includes server and storage technologies and common tools to deliver workload and platform management across your IT environment.

Over the past decade, the number of department servers and department storage has proliferated, creating an IT management challenge. The top driver for consolidation is reduced cost, followed by improved system performance, ease of management, high availability, security, and disaster recovery. In addition to consolidation, many enterprises are interested in managing the growth of their IT resources to maximize return on investment (ROI). To do this effectively, they require usage information from all resources on the network (storage, server, network, and application) to build a complete picture. This information can then be used by the IT staff for optimizing their use of existing resources, improving their service level (through better performance and availability), and proactively managing their capacity planning activities.

On the other hand, users are facing some uncertainties to implement such solution. Lack of skills to realize such virtualization concept or the inability to qualify the value are the main inhibitors amongst implementation. The biggest problem for these services is to know the delivery mechanism – how it is used, how it is charged on basis of usage, etc. Automated processes for cost reallocation and analysis of security and misuse would result in a high level of cost savings.

IT services are viewed as critical to the business. Increases in the number of users, demands for new technologies and complexities of client – server systems frequently cause IT service costs to grow faster than others. As a result, organizations are often unable or unwilling to justify expenditures to improve services or develop new ones, and IT services may become viewed as high-cost or inflexible.

IT Accounting can be used to determine the exact costs of resource usage down to CPU, filestore, and bandwidth, but it is rarely advisable to use this as the basis for charging as the costs of so doing may outweigh the benefits. It is in the interest of all parties to minimize the overall cost of service and bureaucracy, even at the expense of complete precision.

Current leading practice uses IT Accounting to aid investment and renewal decisions and to identify inefficiencies or poor value. A fixed amount is charged for a set capacity determined by the level cf service detailed in the SLAs.

To provide cloud business with a clear understanding of the value they receive from IT, the cost model must be:

- **Equitable:** The chargeback approach must allocate costs proportional to each unit's true consumption of IT services.
- **Controllable:** Business units should have a degree of control over and input into IT spending decisions.
- **Repeatable and predictable:** Charges for a given service should be consistent over a six- to twelve-month period enabling a business unit to forecast its IT costs over the period.
- **Simple:** The chargeback algorithm should be easy to understand, implement, and administer to minimize confusion and overhead expenses.
- **Comprehensive:** All IT costs must be associated with a service. There should be no 'tax' or overhead bucket to account for infrastructure.

A cost model works best when customers understand the pricing structures and their limitations, have some control or influence over the consumption of IT and thus their cost for IT, and believe that the value is reasonable and equitable.

6.1.1 *Service-Based Model*

Recently, there has been a strong push for IT to invoice business units for services described in business terms instead of IT terms. This service-based approach has been driven in part by cost transparency and cost reduction requirements.

The success of a service-based model depends largely on business managers and IT managers working together to define the Service Portfolio, which includes the services the IT organization provides and the cost of these services to the business units. Making IT services understandable to business managers gives them a clear window into infrastructure and application reinvestment. A business director may be persuaded to fund or support infrastructure changes that will drop or increase the consumption or price of services in order to better meet business need.

6.2 RESILIENCY

Resiliency is the capacity to rapidly adapt and respond to risks, as well as opportunities. This maintains continuous business operations that support growth and operate in potentially adverse conditions. The reach and range step of the assessment process examines business-driven, data-driven, and event-driven risks. The goal is to understand the risks to the company, the business process, perhaps the building – whatever concerns you and your business. We may break this step down into detailed examinations because risks in one building, for example, are going to be different from risks in another building. Risks in one geography are different from other locations.

So we will be looking across different parts of the company. We like to focus on one specific area first – maybe a specific business process. By doing so, we usually arrive at the 80/20 rule which says that about 80 percent of issues are going to be common across all business processes, all business entities, and all buildings.

When you use the resilience framework to look at different parts of the company, you are trying to understand whether you have a risk that you can accept or whether you have a risk that you want to avoid and mitigate. In other words, you may choose to do nothing about a risk, or you may improve your infrastructure to help ensure that you can handle events if they occur.

You may also decide that the risk is one that you would prefer to transfer to somebody else, such as business continuity and resiliency services. A lot of organizations feel more comfortable transferring risks associated with business continuity to cloud vendors rather than handling risks themselves, as recovery centres are designed to be robust and to ensure resilience in the face of a disruption.

Additionally, transferring the risk can be accomplished through managed security or resiliency services. This allows you to concentrate on strategic initiatives and leaves the day-to-day management and monitoring of your availability and security configurations to staff locations.

So, what can we recommend to create a framework of resiliency? The resiliency blueprint includes different layers – facilities, technology, applications and data, processes (both IT and business), organization, and finally, strategy and vision.

The framework enables us to examine the business, understand what areas of vulnerability you might have come across – business-driven, data-driven, and event-driven risks – and quickly pinpoint areas of concern and help you understand what actions you can take to reduce the risks associated with those areas.

6.2.1 *Resiliency Capabilities*

The strategy combines multiple parts to mitigate risks and improve business resilience.

- From a facilities perspective, you may want to implement power protection.
- From a security perspective – to protect your applications and data – you may want to implement a biometrics solution. You might want to implement mirroring, remote backup, identity management, e-mail filtering, or e-mail archiving.
- From a process perspective, you may implement identification and documentation of your most critical business processes; you may split functions of processes. You may also want to implement specific requirements confirming to government regulations and standards.

- From an organizational perspective, you may want to take an approach that addresses the geographic diversity, backup of workstation data. You may want to implement a virtual workplace environment.
- From a strategy and vision perspective, you would want to look at the kind of crisis-management process you should have in place. You may also want to examine how you can clearly articulate your security policies to everybody and how you implement change management.

Resilience tiers can be defined as a common set of infrastructure services that are delivered to meet a corresponding set of business availability expectations. Criteria describing resilience tiers were developed by the lines of business and include characteristics/attributes for business impact (for example, revenue), risks (for example, legal), application availability (for example, 24×7), and agility (for example, multiple physical instances).

6.3 PROVISIONING

Provisioning process is a service that uses a group of compliant processes called 'Solution Realization'. Environment provisioning roles separate preparation tasks and assurance tasks from provisioning tasks. Provisioning design decouples provisioning build and integration activities from requirements, design, procurement, and hardware setup. The process formalizes quality assurance testing in preparation for turning over the provisioned product to the customer. Provisioning is a broad-based service that begins with a Request for Service (RfS) to build a fully provisioned environment for the purpose of hosting an application, database, etc. Provisioning can also be invoked when a major modification must be made to the existing environment. Provisioned environments include development, test, Quality Assurance (QA), production, Disaster Recovery (DR). Provisioning defines and communicates what information is required to begin provisioning. The output from provisioning is an environment configured and tested with an appropriate hardware platform, storage, network, operating system, middleware, other system software, backup capability, monitoring capability, and with the application installed per requirements.

- Provisioned products are servers built with all the software and infrastructure required to support a business application.
- Standard solutions are defined so that standard workflows can be derived.
- Design is completed with due diligence before the Request for Service is accepted, including documentation of all specifications.
- Server hardware is assembled, cabled, and connected to the network and SAN before work orders are released to provisioners.

6.3.1 Characteristics

There is an owner providing technical oversight for the lifecycle of each project lifecycle defined as from the initial request for comments (RFCs) through to delivery to the customer. Specifications are reviewed for completeness and accuracy before work orders are released to provisioners. Missing and incorrect information is resolved before provisioning begins.

Provisioner roles for each part of the stack perform build, installation, configuration, and interim verification activities (no change). A status of 'Hold' with a reason code indicates when work orders and the request for service itself are stopped awaiting a response from an external process. The provisioned product is tested, assured for quality, and signed off by the technical owner before being turned over to the customer.

6.3.2 Approach

The environment provisioning process takes an assembly line approach to building a server and integrating its components. To prevent interruption of provisioning tasks due to unforeseen or redundant work, the process defines that upstream activities be completed and signed off before starting downstream activities. Following are the activities discussed:

- Planning precedes execution.
- Validating build specifications precedes building.
- Packaged software installation procedures being tried and tested precedes installing the package on a server.
- Having servers racked, stacked, cabled, and connected to storage and network precedes issuing work orders for provisioning the operating system and base software image.

Measuring achievement is easier without having to account for stops and starts caused by handoffs. It becomes possible to automate building the stack and integrating more components with provisioning tools.

6.3.3 Benefits

This section discusses the benefits of provisioning.

- **Ability to measure progress of all the work related to one RFCs:**
  - Supports the ability to deliver to service levels.
- **Continuous improvement activities based on process measurements:**
  - Enables eliminating delays and learning to continuously provision servers rapidly to shorten the time to deliver.
- **Isolation of the build, install, configure, and customize tasks from requirements, design, and hardware setup activities:**
  - Provides focus for leveraging provisioning automation tools.
- **Role players performing a finite set of repeatable activities:**
  - Enables the collection of intellectual capital necessary for beginning to automate their activities and for planning full automation.
- **An assembly line approach to provisioning:**
  - Facilitates automation of piece parts of the process in an incremental approach to self-service.

Long-term Goals

- Achieve operational efficiencies by using a common set of processes and procedures to deliver provisioning services to the enterprise.
- Achieve target environmental defect rate.

- Establish and achieve Service Level Objectives for delivery of provisioned environments.
- Reduce time to set up development and test environments.
- Reduce hardware/software spending through optimization of all environments and reuse of assets.
- Enforce enterprise provisioning standards.

Short-term Objectives

- Reduce the defect rate for the set up of the development and test environments.
- Improve and provide consistency in the provisioning of environments for all platforms.
- Transfer skills and knowledge of new standard processes and procedures to provisioning teams.
- Gain stakeholder agreement before deployment of a provisioned product that all requirements have been met.
- Reduce rework.
- Improve quality of work experience for process participants.

6.4 ASSET MANAGEMENT

Asset management and change management interact frequently. Several of the activities required to provision an environment rely on RFCs in order to get approval to change known configurations of infrastructure and software components. There are different factors that help to develop the asset management strategy:

- ***Software Packaging:*** Asset management relies on software packaging. The output from software packaging will be used on a daily basis during the installation and configuration of the various software packages requested by the customers. Asset management will only engage software packaging directly when there is an exception. It will pass information so that new or modified packages can be built to enable provisioning.
- ***Incident Management (IM):*** It is used to track any interruptions or issues to the asset management service. These are most likely to be encountered during the OS or application installation, or during the verification of other provisioned components. IM will also be used as an entry point to Problem Management, which will not be engaged by the Asset Management directly. IM's 'business as usual' escalation of recurring incidents as potential problems will contribute to resolving problems related to asset management.
- ***Pool Management:*** Pool management works with asset management to make sure that the products requested are available on the requested date and for the specified duration. Pool management serves as the intermediary process between asset management and the Infrastructure On Demand (IOD) process and activities.
- ***Release Management (RelM):*** It controls the scheduling and testing of additions and updates to environments.
- ***Configuration Management:*** It helps in the absence of a process with its own repository for assets and inventory items.

- ***Systems Management (SysM):*** It is both a process and a service. In order to interface with asset management, it provides all of the information on what attributes of OS, middleware, and business application components need to be monitored. A mature SysM process determines triggers, thresholds, event generation, severity, event correlation automated response, and the tools that will be used.
- ***Operational Readiness Management:*** Asset management interacts with Operational Readiness (OPR) much as other projects and services do. To prepare for release into an environment it is necessary that the documentation describing and supporting the provisioned product align with enterprise standards.
- ***Backup Management:*** EPM links to backup management after the new server is added to the backup script, along with any customizations to the backup job.

6.5 CLOUD GOVERNANCE

One of the major components of any governance model is the proper definition of roles and responsibilities within an appropriate organizational structure. The domain owners within the organization own and are accountable for the business functionality within their proper business domain. These domain owners report to the head, but they also have direct reporting responsibilities within their business domain. These technical roles along with the domain owners strive to achieve a confluence between business and IT. One of the major aspects of cloud governance is to ensure that the lifecycle of services maximizes the value of SOA to the business. In order for governance to be effective, all aspects of the service lifecycle need to be properly handled.

The process transcends all phases of the service lifecycle: model, assemble, deploy, and manage. Each task is numbered based on the phase it falls under.

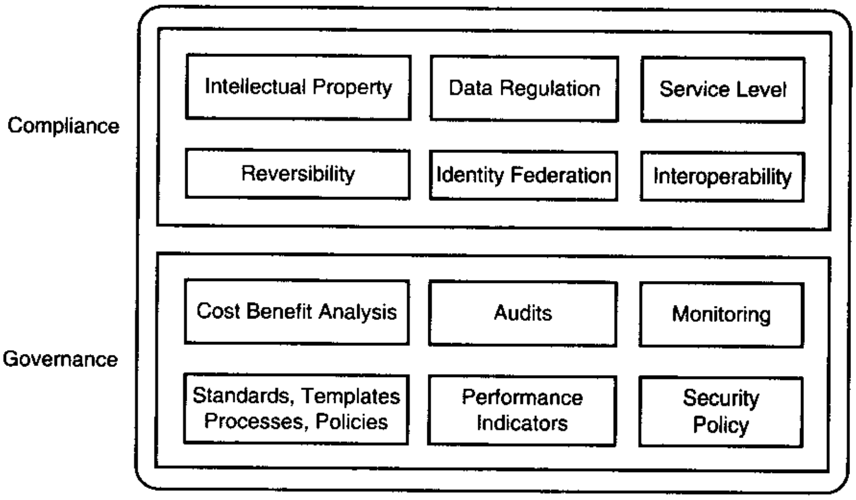

FIGURE 6.1 Compliance and Governance.

The cloud governance scenario should been broken down into realizations (see Figure 6.1). They can be:

- Regulation of new service creation.
- Getting more reuse of services.
- Enforcing standards and best practices.
- Service change management and service version control.

6.6 HIGH AVAILABILITY AND DISASTER RECOVERY

High Availability (HA) and Disaster Recovery (DR) are some of the important factors for cloud deployments. As cloud is based on service models, so different SLAs govern the service-based models to avail the service. HA and DR go together and define the factors for different SLAs between vendor and subscriber to ensure service availability, trust, and helps develop credibility for the cloud vendor. Availability is not just a technology issue – it is a business issue as well. It is sometimes easy for executive management to take infrastructure availability for granted – 'When it's working, you don't know it's there, so it's easy for top management to assume it always will be' – by the business executives, not just the IT executives. The business must be able to make IT investment decisions based on business value. Achieving very high levels of availability usually requires substantial investment (and not just in technology). IT must manage the infrastructure to deliver the required and funded level of availability. But 'the business' must determine the level of availability required to support its business objectives and make the appropriate investments to support that level of availability.

HA and DR have often been treated as separate disciplines, but are converging. HA traditionally focused on avoiding/recovering from non-catastrophic disruptions – server failures, software failures, power failures, network disruptions, denial of service attacks, viruses/ worms, etc. – often relatively short in duration (minutes or hours). It may involve moving workload (dynamically) to another location, but typically does NOT involve moving people. DR traditionally focused on planning for and recovering business operations following catastrophic disruptions.

- Site/facility destruction, hurricanes, tornadoes, floods, fire, etc.
- Often long duration (days to weeks).
- Often involves shifting work (and people) to alternate facilities for some period of time.

Similar disciplines are required for both, but with different emphases. Availability is ability of a component or service to perform its required function at a stated instant or over a stated period of time. It is usually expressed as the availability ratio, that is, the proportion of time that the service is actually available for use by the customers within the agreed service hours. There are different terms to work on:

- **Mean Time Between Failures (MTBF):**
  - The mean (average) time between successive failures of a given component, sub-system, or system.

- **Mean Time To Recover (MTTR):**
  - The mean (average) time that it takes to recover a component, sub-system, or system.
- **High Availability (HA):**
  - The characteristic of a system that delivers an acceptable or agreed-upon high level of service to end-users during scheduled periods. Typically at least 99 percent or more.
- **Continuous Operations (CO):**
  - The characteristic of a system that allows an 'end-user' to access the system at any time of the day on any day of the year (24×7×365).
- **Continuous Availability (CA):**
  - The characteristic of a system that delivers an acceptable or agreed-to high level of service at any time of the day on any day of the year (24×7×365).
- **Availability Management:**
  - The process of managing IT resources (people and technology) to ensure committed levels of service are achieved to meet the agreed upon needs of the business.

Recovery capability is the process of planning for and implementing expanded operations to address less time-sensitive business operations immediately following an interruption or disaster. Recovery Time Objective (RTO) is the period of time within which systems, applications, or functions must be recovered after an outage (say one business day). RTOs are often used as the basis for the development of recovery strategies, and as a determinant as to whether or not to implement the recovery strategies during a disaster situation. Recovery Point Objective is the point in time to which systems and data must be recovered after an outage (for example, end of previous day's processing). RPOs are often used as the basis for the development of backup strategies, and as a determinant of the amount of data that may need to be recreated after the systems or functions have been recovered.

Disaster Recovery is the process of creating, verifying, and maintaining an IT continuity plan that is to be executed to restore service in the event of a disaster. The objective of the Disaster Recovery Plan is to provide for the resumption of all critical IT services within a stated period of time following the declaration of a disaster.

- Protect and maintain currency of vital records.
- Select a site or vendor that is capable of supporting the requirements of the critical application workload.
- Provide a provision for the restoration of all IT services when possible.

A Disaster Recovery Plan includes procedures that will ensure the optimum availability of the critical business functions and the protection of vital records necessary to restore all service to normal. The Disaster Recovery Plan is dependent upon and uses many of the same recovery procedures as those defined and developed by the Recovery Management process. The execution of the Disaster Recovery Plan will use many of the same policies, procedures, and staff as defined in the Crisis Management process. The DR event is primarily a 'crisis' of greater magnitude and scope than the situations that are routinely managed on a day to day basis.

The true business need for high availability of IT systems including rapid recovery for disaster situations must be determined and justified. The cost of down time must be understood by business unit to establish true business need for HA and rapid DR. An availability strategy is required to guide the organization in implementing high availability and support rapid recovery from a disaster.

- Align the IT strategy with the business strategy and requirements.
- Justify investment in HA and DR initiatives.
- Ingrain HA in the IT culture.
- Define a robust IT architecture and invest in building HA into the design of the infrastructure.

When disaster recovery plans fail, the failures primarily result from lack of high availability planning, preparation, and maintenance prior to occurrence of the disaster. Lack of an IT architecture employing hot back-up components and hot back-up sites inhibits achievement of Continuous Availability across component failures or site failures resulting from disaster. Recovery severely delays when many back-up components have to be rebuilt from scratch. Change Management processes fail to ensure backup components and recovery documents are updated simultaneously with primary component upgrades.

The technology must fully exploit high availability design techniques such as redundancy with hot back-up capabilities to support rapid recovery.

Application and data interdependencies are important considerations in determining business function priorities. Network connectivity must consider more than connectivity between the datacenter recovery site and the user site. Consider connectivity to business users, system to system, customers, and outside agencies. Consider failure of multiple sites and setup for connectivity from back-up site to back-up site. Where critical business unit users must support the recovery effort, for example to prepare for end of day processing, immediate access to workstations is required. If the business processes are dependent on printing, printer recovery must be treated with appropriate priority.

The events of previous disasters confirm that effective and rapid recovery from any disaster is dependent on mature processes supporting high availability. Service level requirements and business requirements must be understood and objectives negotiated. An infrastructure supporting high availability is essential to rapid disaster recovery. The system and application designs must be built to support high availability and rapid disaster recovery. Complete configuration information is necessary to reconstruct all system platforms following a disaster. Adequate testing must validate the capability of the plans and ability to perform the procedures, whether for day-to-day high availability or for disaster.

To prevent gaps in disaster recovery plans, recovery procedures, technology platforms, and DR vendor contracts must be updated concurrently with changes. Fast and effective recovery from a component outage or from a disaster requires well thought out, pre-developed, tested, documented, and practiced recovery. Defects and shortcomings must be resolved quickly to ensure the plan will work.

6.7 CHARGING MODELS, USAGE REPORTING, BILLING, AND METERING

Today, enterprise business units' budgets fund 60 to 70 percent of Central IT's services. The other 30 to 40 percent is funded by other means, so it is clear that in general organizations do not use a single charging mechanism, but a combination of mechanisms for different purpose to achieve an overall solution. Existing processes were institutionalized in many large organizations decades ago and the responsibility has been passed down from employee to employee over generations. The pitfalls of chargeback are well-documented; they include user architectural rebellion, IT investment vacillation, bureaucratic excess, and malicious obedience to IT standards. These pitfalls fall into an IT-centric view of providing service; these arguments and others like them seem shallow when presented with the business imperatives that are often at the root of maintaining a chargeback system.

6.7.1 Challenges

Many organizations do not implement sophisticated internal chargeback mechanisms due to the complexity. You have to be able to determine all the metrics, and be able to break them out by user; you have to keep track of what organizations the users are in, which is not a simple task. This creates a large volume of data for the items that you can directly tie to a user (for example, CPU, memory, and disk that are associated with a particular transaction). While this is reasonably easy to do in a dedicated workload environment, enterprise environments add another layer of complexity.

Then you have to add in the overhead (operating system, program products, network, support, and processes such as space management). When you introduce the allocating of details such as, for example, the cost for SAN ports in a switch, you may have more of a political discussion than a technical one, as you may not be able to tie back specific items to transactions.

6.7.2 Benefits

The benefits from implementing a more effective system can be enormous. The following are the advantages for managers looking for the benefits of implementing a more comprehensive chargeback system. When forced to confront the issues of chargeback implementation or chargeback system changes, managers should align their practices with the benefits of a chargeback system.

Charging for services will not solve all the problem of IT department, nor will it be the source of all service problems when dealing with business managers. IT managers must leverage a chargeback system to harvest opportunities for improving and streamlining service delivery.

6.7.3 Cloud Chargeback Models

In consolidated environments, IT custodial service employs a cost recovery mechanism called chargeback. Chargeback is a mechanism to institute a fee-for-cloud-service type of model. Chargeback allows for IT custodial to position their cloud services as a value-added service, and use cost recovery mechanism to provide varying degrees of cloud service levels, at differentiating costs. To device an effective chargeback model, it is imperative that the IT organizations have a complete understanding of their own cost structure and cost breakdown by components used as resources. The clear understanding of costs is important in devising a chargeback mechanism and a utility like model to justify the billing costs associated with use of various resources (Figure 6.2).

When it comes to employing chargeback models there is no silver bullet that will solve the perceptions and all of the user expectations from a cloud services commodity model. There are various models prescribed and practiced in the industry today, and each of these models will have to be evaluated to see which one best fits the cultural and operational boundary of the organization. Here we discuss a few chargeback models. An organization may adopt a 'hybrid' model and combine the feature of more than one model.

- **Standard Subscription-Based Model:** This is the simplest of all types of model. This model entails dividing the total operational costs of IT organization by the total number of applications hosted by the environment. This type of cost recovery is simple to calculate, and due to its appeal of simplicity, it finds its way in many organizations. The year-to-year increase in IT costs due to growth and expansion is simply added to the costs of subscribers tab. While this is a simple chargeback model, it is fundamentally flawed, as it promotes subsidy and unequal allocation of resources. So, with this model, a poorly performing application is subsidized by other applications, also less emphasis is paid to resource consumption and application footprint.

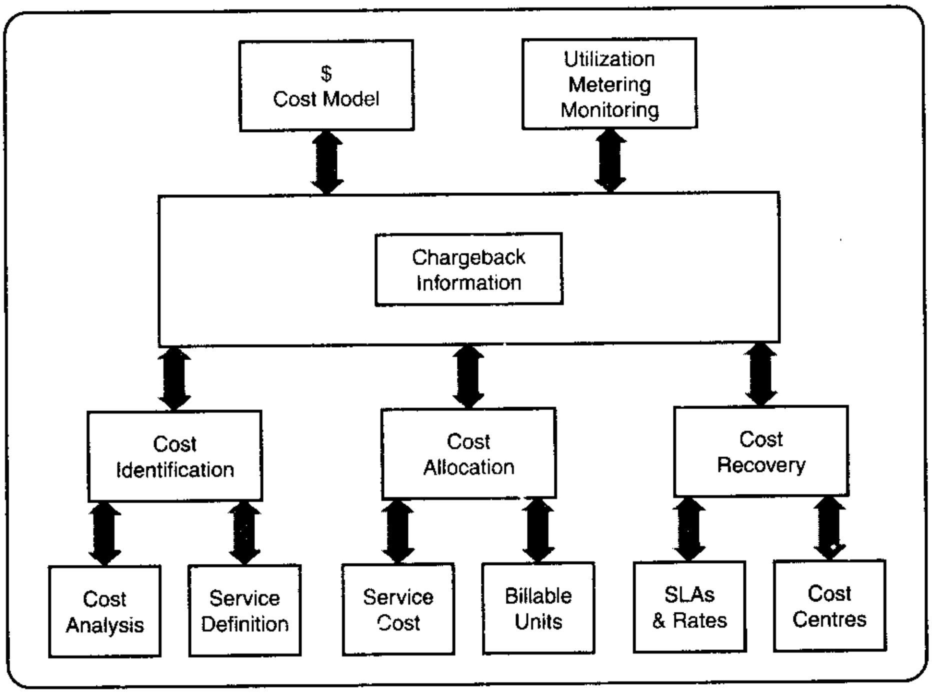

FIGURE 6.2 Chargeback model.

- **Pay-Per-Use Model:** This model is targeted for environments with line of businesses (LOBs) of various sizes, and unlike the standard subscription model, this model emphasizes on charging based on application's consumption of resources and choice of service level agreements (SLAs). So, for instance, a poorly written application may pay more for shared services simply because of its footprint, or an application's desire for higher degree of preference or dedicated resources would pay more due to choice of service policy. This model can be complicated in its approach, simply due to the framework around resource usage and its monitoring. While this model ensures fair and equitable cost recovery, it may take longer to arrive at agreeable metrics and cost models associated with resource consumption.
- **Premium Pricing Model:** The premium pricing model focuses on class of service and guaranteed availability of resourcesfor business applications. As the name suggests, the LOBs will incur a premium for preferential treatment of application requests, and priority in resource allocation, during times of contention to meet the service goals fully. This can also include dedicated set of hardware nodes to host applications. So depending on degree to isolation and separation from the shared services model, the price tag can go up. Such types of models are usually preferred by LOBs with mission critical and high impact on revenue applications. Also, this model will usually never exist alone, and may coexist with other base line chargeback models such as standard subscription based or pay per use model.
- **'Hybrid' Model:** The 'Hybrid' model attempts to adopt best of breed models and offers the combined advantages of two or more chargeback (CB) models. For instance, a CB model can be devised which has a flat entry fee per application to use the infrastructure in addition to paying for consumed resources. So this way the flat fee can be characterized as a fixed expense to be hosted, and the additional costs for resource consumption can be linked to variable cost. This hybrid model, for example, combines the standard subscription model and pay per use model, and takes advantage of utility like billing service, much like utilities at home. Since there is no one single CB model that will fit all environments, it is not uncommon to see a combination of one or more types of models. While there are advantages of using one single chargeback model for entire organization, there may be value in introducing catalogue of CB models proving a choice to LOB to pick from. This will not only provide flexibility to IT organizations and LOBs alike, but also present LOBs with many choices to pick from. With this type of flexibility, there will be no need for one agreeable model, as each LOB will have an array of choices to pick from. Catalogue of chargeback models will only work if every model has been tested for equality and value they provide. For instance, one application with all things constant, should have same cost of operation under every model, and the costs should vary with volume, QoS (Quality of Service), and chosen service policy, and this is where the choice of model will add value to the LOB applications.

Simplifying Chargeback

Chargeback is an important tool to align IT services with business objectives; it is also a tool that can bring together IT and business operating community to understand the value created by shared environment services.

The ultimate goal of any chargeback model in a shared environment is to provide the business with resilient and robust value-add IT services at a competitive costs. These cost advantages are enabled by efficient use of hardware and software resources, and the resiliency comes from harnessing the computing power of virtualized grid like IT infrastructure.

Chargeback models, by the very nature of the intent, can be complex and may require extensive education to business and financial community alike. The other challenges include IT organization and business community to agree upon the value and the language around cost associated with value.

Simplifying chargeback is vital to adoption and acceptance to a shared service infrastructure. First step to simplification is education on purpose and intent of adopting the education can also be used to gather RFCs on appropriate chargeback models. This will provide a baseline of mindset around chargeback. The second step should include a complete breakdown of IT organization's costs. This transparency will encourage better understanding of operational costs by other participating LOBs. The next natural step would be to device a model that is agreeable to all. This is where the RFC from the first step would be instrumental in working on a common model; which resonates with accepted financial practices. It is recommended using more than one model; in the initial phase, this will allow for review and analysis of other models, encouraging LOB participation. After complete review of all the used models, the best model should be chosen, that may reflect the best choice that presents with an equitable and fair chargeback mechanism. The overall organization's operational and financial reporting practices will also play an important role in determining the choice of model to be accepted. This process of simplifying chargeback will ensure participation from all business units, and bring forth the (operational and financial) constraints from the inception, thereby resulting in a universally accepted chargeback practice.

6.7.4 IT Infrastructure Governance

Governance in a shared infrastructure becomes paramount, as resources shared by all business units require some level of policies and control mechanisms that define the boundaries and upholds the business unit requirements. The isolation preference enables infrastructure to dedicate nodes to a specific application or group of applications. With such a requirement, there ought to be governance to ensure that these requirements are adhered to. Chargeback models may reflect the higher costs associated with such a requirement, but governance ensures that the cost allocation is fair. A sound governance policy will ease change management and institute higher confidence in shared infrastructure services.

There can be many features that are instrumental in proving a flexible virtualized run-time environment for all of the enterprise applications, with some features specifically enhancing productivity in manageability of the environment. These features, such as application editioning, chargeback, unified administration, and so on, will require a thorough review of existing practices or provision the inclusion of new practice. Governance of such an environment will include ownership, accountability, and access to operational environment. IT infrastructure may be too broad and may include all aspects of IT operational management and control. The key to successful adoption is to begin with categorizing the features into existing governance models. This way the clear separation of responsibility is maintained and any change can be easily absorbed. Moving forward, new tasks and models may have to be introduced (for example, chargeback and service policy governance) to add to overall IT operational control. Like any change management practice, this process should be expected to be a long-term undertaking, even to the extent of projectizing the effort with active participation from upper management.

6.7.5 Basic Requirements

The area of business unit contribution to IT funding can cause significant friction between business units and IT managers. For this reason the business units need a documented understanding of what they are getting (that is, value) for their money. The chargeback metrics used in determining the individual business units' share of funding contribution should be directly tied to the Service Level Agreement between central IT and the business unit and should reflect the following elements:

- **Fairness:** The chargeback approach must be seen as allocating costs in proportions that reflect each unit's true consumption of information and communication services. One group should not be subsidizing IT usage at the expense of another.
- **Control:** Business units should have a degree of control over, or input into, IT spending decisions.
- **Repeatability and Predictability:** Charges should be repeatable (there should be consistency in the application of data collection and charging methods so that charges are consistent) and predictable (a business unit should be able to create a reasonable forecast of its expected charges over a six to twelve month period).
- **Simplicity:** The chargeback algorithm needs to be easy to understand and simple and inexpensive to implement.

Chargeback works best when customers understand the pricing structures and their limitations, have some control or influence over costs, and believe that the policy is reasonably fair, given the limitations of a particular system and the underlying business rationale behind chargeback.

Chargeback Schemes

Possible chargeback approaches are listed below. It may well be that the best solution is to use a combination of these for different aspects of the IT infrastructure.

Allocation-Based

In this model, IT service costs are buried in corporate overhead as a budget line item, usually determined one year at a time. The model has nothing to do with usage; instead, it charges business communities based on their position within the enterprise (for example, the share of employees, unit shipment volume or total revenue). This model is attractive as it is the simplest and costs least to implement. However, some weaknesses are:

- The difficulty of rebalancing the scale when the business measures change.
- The lack of incentive for end-users to control their resource usage.
- The frustration of business managers unable to control or influence their budget share – although at least it will be predictable.

Flat Fee

This model adds elements of negotiation and capacity planning. The IT organization determines what percentage of the IT service workload a business area represents, calculates a preliminary package rate for that area, then negotiates a rate with business managers. For example, if finance represents 8 percent of the IT workload, it might pay a proportional fee. Because the flat fee is tied to usage, it gives business managers a chance to understand what they are paying for. Flat fee is appropriate for environments in which third-party application packages are used heavily. Variations such as access fees and subscription fees can be relevant to certain components of the system such as the network and specific end-user services.

Resource- or Usage-Based (Direct Cost Recovery)

Resource-based costing and its most common form of implementation, usage-based costing, focuses on developing a standard unit cost for each major resource type or category that best represents the use of that resource. For example, the measure for CPU usage could be CPU seconds consumed by an application, for storage usage it could be number of Gigabytes of storage occupied by an application or business, for the network it could be number of bytes transferred. The basic idea is that the costing unit represents some measure of the resource consumed that can be traced back to the user of that resource.

This method requires that all elements of the IT infrastructure and associated software specific to the application be identified and are directly charged to the end-user on a per-user basis. The cost per unit (whatever unit is chosen) needs to cover all IT-related costs – operations, support, buildings, networks, etc. There may be parts of the 'enabling infrastructure' that are chosen to be recovered through other methods such as allocation, flat fee, or a per user charge.

This cost-recovery model is still widely used as a traditional mainframe approach. However bundling mainframe computing services into resource-based charges can create bloated CPU fees, which prompt users to purchase their own systems. This approach is not always effective in the complex PC-based and distributed computing environments where the mechanics and time involved in tracking usage may cost more than the IT organization recovers. Moreover, the language in a resource-based chargeback scheme is so techno-centric that the bill mystifies business managers.

Product-or Service-Based

In the commercial environment, there has been a strong push for IT to invoice the lines of business or business units in business terms instead of IT terms. This means that instead of charging a business unit for CPU seconds consumed (or in the case of networks, the number of bytes transferred), this model defines IT costs in measurable events, transactions, and functions that are relevant to the business and outside the IT organization, for example, invoices produced, cheques written, e-mail messages sent, reports delivered, number of claims processed, number of policies written, or some other metric that represents the work performed. This method could be called a business product-based approach whereas the resource-based method is an IT product-based approach (say, bills are expressed in IT terms like CPU seconds).

In any case, the product-based approach requires all the data collection instrumentation and methods used in a usage-based approach to be in place and then expanded to include the mapping of that usage data to product and service categories in addition to department categories.

The success of the model depends on business managers and IT organizations together defining specific services that the IS organization agrees to provide and for which the business agrees to pay. Making IT services understandable to managers gives them a clear window into infrastructure and application reinvestment.

Activity-Based

Activity-based costing (ABC) is the most difficult of all the methods to develop and implement. There are almost no large IT organizations that have a truly activity-based approach to IT costing. The IT organizations that claim to have activity-based costing usually don't – they have product-based costing. Quite a few organizations have started activity-based costing efforts within IT but stop them before completion and settle them product-based costing.

Activity-based costing assigns costs to each activity that goes into delivering a product or service. ABC is a cost methodology which:

- Derives the costs of an organization's outputs (products and services).
- Identifies the activities and tasks (processes) used in the production and delivery of the outputs.
- Identifies the resources consumed in the performance of these processes and instruments these activities so that the cost per task can be rolled up into a charge per major activity by department.

ABC takes resources (that is, expenses from the general ledger accounting system), moves them to activities (that is, moves those costs to activities), and then moves the costs from activities to cost objects (that is, product and services).

Activity-based management (ABM) is the method or process of using the data produced by ABC. So ABC produces cost information, ABM takes that information and uses it to find ways to improve those costs and the overall operations of the organization.

While IT product-based costing produces a charge by product (for example, claim or policy), IT activity-based costing produces a charge by activity (for example, printing a claim or handling a policy, or printing a report). As can be seen by this example, the level of detail required and reported is significantly higher than product-based costing.

Activity-based costing is the 'premiere' approach of cost accounting options. Its strength is that costs can be managed very well since each activity has a cost driver that can be measured. So, a charge area such as 'handling a claim' may be broken down into 10 IT service activities. These activities can be ranked by their total contribution to the overall cost of 'handling a claim'. This allows cost managers to focus their time on the largest contributors to cost by activity.

Even though this information is extremely valuable to decision-makers, the cost of getting this information can be prohibitive for all but the most disciplined of institutions. It requires a major investment of time, people, and resources to build and maintain an ABC system, with an extended implementation period.

External Pricing ModelIMarket-Based

This model is geared toward turning a profit. The model presumes that an IT organization operates in a fairly open market inside and outside the enterprise and requires some 'market testing' for the cost of services. Pricing is determined by what is available on the outside market. Although a small percentage of midsize enterprises use this model, it has clear advantages; it may well become a hallmark of IT organizations recognized as value-generating service providers.

Determining the correct pricing can be expensive if survey is required to support the scale of charges. However, basic comparison with contract, staff, and consulting rates and high-level assessment of IT spends against benchmarks is sufficient to support the cost model.

Very often, profit centre cost models will lead to over recovery. This can be corrected with a simple adjustment within the accounting process. However, care must be taken not to under recover costs. Year-end upwards corrections will often cause friction with the business units. Most organizations look to over recover by a small percentage for Cost Centre accounting.

6.8 SUMMARY

Today IT delivers technology to the business units and assesses charges based on the number of devices provided. The business units do not have the ability to identify the elements of this cost and cannot manage their consumption of the technology.

This chapter addresses this problem by bundling the technology IT offers into services for the business units to purchase as needed. The Cost Model Strategy detailed in this chapter provides recommendations about how to design and implement an equitable, accurate, and auditable method of charging for services that provide value to customers.

CHAPTER 7

CLOUD VIRTUALIZATION TECHNOLOGY

7.1 INTRODUCTION

Today, cloud is the buzz word in the industry. The advent of powerful virtualization technology in the infrastructure domain gives us the options to reap the benefits of the cloud deployments. The powerful line-up of servers blended with advance Web technologies gives ease to exploit the powerful features of virtualization combined with cloud concepts. Continuous improvement is leveraging technology and expertise to do the same things more efficiently. Continuous innovation is the fusion of new business designs and next-generation technologies to actually do things differently, not just once, but over and over again. While virtualization sounds like a very complex, technical thought, it's really a simple idea.

Virtualization is a fast-growing infrastructure in the IT industry. New technologies are being introduced. As a result, technology providers and user communities have introduced new set of terms to describe the technologies and their features for virtualization. Some of the terms overlap with the others and some may be ambiguous. For example, 'Hardware-Assisted Virtualization' and 'Hardware-Based Virtual Machine' refer to the same thing, while the term 'Paravirtualization' may be something new for some users.

Virtualization represents the logical view of data representation – the power to compute in virtualized environment, storing the data at different geographies and various computing resources. This removes the restrictions on computing like difficult infrastructure deployments, collocated computing resources, physical movement, and packaging of resources.

This statement really makes two points:

1. To virtualize your systems, you separate the physical from logical, you manage and utilize IT resources as a cohesive, holistic unit that is constantly adjusting, reallocating, and responding as changes in the business environment dictate.
2. Virtualization is a liberating technology – meaning you have better, more responsive access to information. You can further simplify IT management by instituting policy-based response, and ultimately you reduce the cost of operations.

For example, in storage, rather than saying we have three different types of storage systems, which together might total 30 TB of disk space – we start managing the 30 TB as a single type of resource. We focus on how to use the resource and not how to manage it.

It is a technique we have been using in large mainframe computer for 30+ years; not having to manage each computer or resource separately – but to manage them together, virtually. This allows for huge improvements in utilization. A typical mainframe today runs at between 70 and 90 percent utilization. The rest of a company's infrastructure is probably running at less than 15 percent utilized. Raising the level of utilization across your whole infrastructure usually means you will need to manage fewer things. Fewer things require fewer people and less infrastructure expense.

By extracting some of your administrative cost out the infrastructure and by increasing system and resource utilization and improving productivity, these 'virtualized' IT assets can help fuel the business growth we just talked about, control the cost in doing so, and increase staff productivity at the same time. Having simplified virtualization implementation as a first step, it is then easy to automate, which means reduced errors and streamlined business responsive systems.

7.2 VIRTUALIZATION DEFINED

Virtualization isn't a vague concept – you probably are already engaged in virtualization in some fashion – but it helps to understand virtualization as a process. So how could a virtualized environment help your organization?

Virtualization is an abstraction layer (hypervisor) that decouples the physical hardware from the operating system to deliver greater IT resource utilization and flexibility.

Virtualization allows multiple virtual machines, with heterogeneous operating systems to run in isolation, side-by-side on the same physical machine.

So, how can virtualization help business? Virtualizing the service infrastructure can provide substantial benefits:

- **Save money:** With virtualization technology, you can reduce the number of your physical servers, and therefore, the ongoing procurement, maintenance, and ongoing operational costs.
- **Dramatically increase control:** Virtualization provides a flexible foundation to provide capacity on demand for your organization. You can quickly deploy new servers and therefore services in minutes as it is easy to ship infrastructure when we deploy it using virtualization techniques.
- **Simplify disaster recovery:** More efficient and cost-effective disaster recovery solution can be realized with virtualization technologies. Imagine bringing your servers and business on-line at an alternate site within minutes. It is possible using virtualization.
- **Business readiness assessment:** Virtualization introduces a shared computing model to your enterprise as it is easy to understand the infrastructure requirements in virtualized environment and there is no need to implement it physically.

Depending on the organizational structure, virtualization change may either impede or enable the virtualization strategy.

7.2.1 Why Virtualization?

Let us now look at the need for virtualization in the infrastructure domain. Virtualization can help you:

- Lower the cost of your existing infrastructure by reducing operation and systems management cost while maintaining needed capacity.
- Reduce the complexity of adding to that infrastructure.
- Gather information and collaboration across the organization to increase both the utilization of information and its effective use.

- Deliver on SLA response times during spikes in production and test scenarios.
- Build heterogeneous infrastructure across the whole organization that are more responsive to the organization's needs.

Being able to implement solid information management solutions in your organization does not mean you have to change your whole IT environment in one major re-engineering project. There is a stepped approach that we see most successful companies follow. Some people may focus more on automation capabilities, other may focus more on virtualization, but it is the breadth of capabilities across the spectrum of information management that truly unlocks the value of your IT infrastructure.

The first steps in the process are to simplify your environment by consolidating like systems platforms onto fewer, more manageable resources. For the past decade, this has been one of the primary ways companies seek to reduce costs and increase utilization.

Once you have brought like systems together into a more efficient structure, you can start to automate the management of those resources, adding and moving capacity as needed. Allowing business needs to drive resource usage rather than resources dictating how well the business performs.

Automation of tasks such as increasing or moving capacity can lead to progression of task automation all associated with a given process or sub-process, such as application testing or release management of an updated production configuration. You may want to look at ITIL for guidance on IT processes, which in turn, can lead to insights on the highest priority tasks and processes to consider for automating.

Another key activity at this point is to start bringing together these consolidated resources across functions within the company. Begin breaking down the silos of technology and sharing resources across functions within the enterprise. By doing this, a company can use resources that may sit idle at various times of the day to perform tasks that are overburdened at those same times. The ability to share these resources in a seamless fashion gives companies the ability to quickly respond to changing business needs without over investing in technology.

In order to use these resources most effectively, companies cannot allow the standard process that may be in place today to slow down the adaptation of resources to new workloads. To facilitate the fluid environment, we need tools and processes that allow the automated orchestration of resources to respond to those business challenges.

As virtualization and automation capabilities improve within an organization, we see companies being able to move to enterprise-wide virtualization that is enabled by a global virtualization fabric. This fabric utilizes advanced virtualization techniques available through grid technologies and more advance mainframe virtualization platforms to allow seamless access to resources wherever they exist within the organization. It begins to eliminate boundaries between resources that have been created by organizational silos or management processes.

Finally, we see organizations using these advanced virtualization concepts not only to access resources within their organization, but being able to truly see resources on demand; whether they are within the company or outside at partner or vendor locations. In this state, resources are available when needed, peak demands can be serviced without keeping unused capacity on the floor for extended periods of time, and information flows seamlessly between organizational functions, both within and outside the company.

One of the most critical ingredients for successful enterprise-wide and inter-enterprise resource sharing, application integration, and business process collaboration is security management. Fundamentals such as authentication, authorization, and access must be in place across systems, networks, and applications. Establishing roles and using identity management will save time and money in the long run, and tighten security immediately. Having the right solutions providers for security and verification between your suppliers, partners, and customers will help you get to that 'always on' state.

All of these infrastructure management techniques are available today, but many companies find it difficult to implement them as rapidly as they would like due to outdated IT governance and management processes. Because of that, companies must address those processes and cultures that hold them back from taking full advantage of the technologies available today.

7.2.2 *Infrastructure Virtualization Evolution*

First of all, the objective is to take what's complex today and to try to do physical consolidation with it. Increasingly physical consolidation is becoming easier to do as the processes deliver even greater virtualization capabilities and are more flexible. But ultimately, there is only so far that you can go with physical consolidation. It is easy to claim that you can migrate all your Window servers to a mainframe Linux-based system, but it is not so easy to do. This has a huge potential for doing consolidation. But it's not so easy to do. It takes time. So one of the key things in terms of what we're doing with virtualization is to treat things much more logically. What we want to do is to get into an environment where the resources that make up your computer systems, local or remote, are one logical pool of resources that you can use as the business applications need. So if you've got smaller servers that are capable of running additional work, it can be done automatically and dynamically rather than trying to get more value from them by physically removing them and taking the workload and putting it on a bigger system. However, the two things are complementary, the ability to deliver logical consolidation and logical simplification, as well as physical consolidation.

Virtualization of physical machine resources has been used in mainframes for production workloads for many years. Different virtual machines can run different operating systems and multiple applications on the same physical computer. Each virtual machine is encapsulated and segregated, and contains a complete system including CPU, memory, and network devices to prevent conflicts and allow a single physical machine to safely run several different operating systems and applications on the same hardware.

7.3 VIRTUALIZATION BENEFITS

Traditional benefits of virtualization include:

- Server consolidation.
- 'Green' IT – reduced power and cooling.
- Reduced hardware costs.

Virtualization benefits have expanded to include:

- Increased availability/business continuity and disaster recovery.
- Maximized hardware resources.
- Reduced administration and labour costs.
- Efficient application and desktop software deployment and maintenance.
- Reduced time for server provisioning.
- Increased security on the desktop client level.
- Dynamic and extensible infrastructure to rapidly address new business requirements.

7.3.1 Current Virtualization Initiatives

We now see what are the new initiatives are active in the industry and how they help us in infrastructure domain:

- **Virtual CPU and Memory:** Physical CPUs and RAM can be dedicated or dynamically allocated to virtual machines. As there is no OS dependency to physical hardware, with CPU checking off, virtual machines can be seamlessly migrated to different hosts with the background changes to physical CPU and memory resources being transparent to the guest OSs running on the virtual machines.
- **Virtual Networking:** This creates a virtual 'network in a box' solution that allows the hypervisor to manage virtual machine network traffic through the physical NIC(s) and allow each of the virtual machines to have a unique identity on the network from the physical host.
- **Virtual Disk:** SAN-based storage is presented as storage targets to the physical host, which in turn, are then used to host the virtual machines' vdisks.
- **Consolidated Management:** Performance and health of the virtual machines and guest OSs can be monitored and 'console' access to all of the servers can be accessed via a single console.
- **Virtual Motion:** Active virtual machines can be seamlessly and transparently migrated across physical hosts with no downtime and no loss of service availability or performance. The virtual machine's execution state, active memory, network identity, and active network connections are preserved across the source and destination hosts so that the guest OS and running applications are unaware of the migration.
- **Storage Virtual Motion:** The vdisks of active virtual machines can be seamlessly and transparently migrated across data stores while the execution state, active memory, and active network connections remain on the same physical host.
- **Dynamic Load-Balancing:** Dynamically load balances virtual machines across the most optimal physical hosts to ensure that pre-defined performance levels are met. Virtual machines can be automatically and seamlessly migrated to a less busy host if a particular host in a resource pool is in a high utilization state. Different resource pools can be defined for different business needs. For instance, production pools can be defined with more stringent service level requirements while development pools can used more relaxed service levels.
- **Logical Partitions (LPARs):** Hardware layer logical partitioning to create two or more isolated computing domains; each with its own CPU, memory address space and I/O

interfaces and each capable of housing a separate operating system environment, on a single physical server. LPARs can share CPUs or have dedicated physical CPUs. Likewise, an LPAR can be dedicated physical memory address space or memory addresses can be dynamically allocated among LPARS as needed.

- **Logical Domains (LDOMs):** The operating systems running in each logical domain can be independently managed, that is, stopped, started, and rebooted, without impacting other LDOMs running on the host. A Type 1 'bare-metal' hypervisor isolates computing environments from physical resources, notably, the separation of domains across distinct threads using the multi-threading technology because the hypervisor is dynamically managing and encapsulating the allocation of physical resources.
- **Zones:** Zone is an operating system-level virtualization solution rather than a hardware-level hypervisor solution. Each zone is an encapsulated virtual server environment running within a single operating system instance. As such, zones share a common kernel, through a global zone, although 'non-native' zones can emulate an OS environment other than that of the host's native OS. Zones allow for virtualization across a single physical server platform, but some applications may still be limited in their ability to run within zones if they require direct manipulation of the kernel or it's memory space (since the kernel is shared across zones) or if the application requires privileges that cannot be granted within a non-global zone.

7.3.2 *Virtualization Technology*

Advances in computing, especially in Hardware technologies, are driving adoption of virtualization and help to meet the corporate demand for computing that has grown exponentially over the past decade. Success of virtualization concepts over a period of time has led to the genesis of better infrastructure optimization. Virtualization gives several benefits like Live Migration, Hardware Support Virtual Machines, Management of Virtual Datacenters, Virtual Networking Performance, Networking Support, Dynamic VM storage, Broad OS Support, Network Load Balancing, etc. These can be realized via a virtualization platform which allows automatic provisioning of environments and deployment of applications into those environments. In addition to setting up a virtualized infrastructure with self-service capabilities for provisioning, scaling, monitoring, and de-provisioning, the solution shall address application environment issues through best practices and automation.

A virtualized environment allows automatic provisioning of environments and deployment of applications. This infrastructure should enable the dynamic and repeatable process to create environments that will result in cost savings in terms of infrastructure costs and manual interventions. This platform capability for allowing back-up of VM images for subsequent environment setup requests should be used for eliminating application deployment and configuration issues. This infrastructure reduces the time required to obtain and boot new server instances, allowing upgradation of scale capacity quickly, both up and down, as computing requirements change. The solution should provide the visibility into resource utilization, operational performance, and overall demand patterns — including metrics such as CPU utilization, disk reads and writes, and network traffic.

The enterprises work load is not constant. The load on the activities can be more during the peak hours and less during other hours. So the computing resources have to be allocated more during the peak hours and vice-versa during off peak hours. Downtime for the datacenter infrastructure, whether planned or unplanned, brings with it considerable costs. It should be ensured for higher levels of availability that have traditionally been very costly, hard to implement, and difficult to manage.

As the datacenter is virtualized, it is capable of delivering uncompromised control over all IT resources with good utilization efficiency of the available resources. Virtualization of any datacenter would help to gain high performance, scalability, and flexibility. These critical underlying factors lay the foundation for adopting hypervisor for the datacenter virtualization and for accelerating this infrastructure of the transition to a virtualized computing model in near future.

For the purpose of virtualization, the section below explains how the various features of virtualizations can be used in the virtualized datacenter. It also acts as the foundation for virtualized computing and it supports various applications that help in virtualization of business critical activities.

Hypervisor

The unique features of hypervisor on bare metal enable the hardware to be used efficiently. There are various memory optimizing techniques that will make the hardware to over-commit on the available resources and also efficiently give high availability to the virtual machines running on the host servers. The power-saving features also enable the datacenters to GO Green and also save energy by powering off the servers. This is done automatically as per the load of the infrastructure environment. The servers would be automatically brought up and running when there is a requirement for more computing resource for processing the load on the infrastructure. Bare-metal hypervisor uses high-level resource management policies to compute a target memory allocation for each virtual machine based on the current infrastructure load and parameter settings for each of the virtual machines. The computed target allocation is used to guide the dynamic adjustment of the memory allocation for each virtual machine in the infrastructure. In the cases where host memory is over-committed, the target allocations are still achieved by invoking several lower-level mechanisms to reclaim memory from virtual machines.

Administration

To administer the virtualized datacenter activities a single console application is required. Using the centralized management software, virtual management server centrally manages bare metal hypervisor environments allowing IT administrator's centralized control over the virtual environment. Administrators can provision VMs and hosts using standardized templates, and ensure compliance with hypervisor host configurations and host and VM patch levels with automated remediation.

This administration software constantly monitors the virtualized datacenter. Also it would allocate the necessary computing resources and de-provision them as and when required. For load balancing, this administration application would also do a dynamic movement of the VM servers from one bare-metal hypervisor server to another without any disruption in the services offered by that respective server.

7.3.3 Virtualization Use Cases

The following section describes the virtualization functionalities that can be used for the datacenter applications and how virtualization improves the functionality in any datacenter environment.

Availability of Machines

This feature makes the machines in the virtualized datacenter as High Available. This would ensure that multiple datacenter activities are carried out even on the event of Hardware failures. This feature should be configured and used for all the virtual machines in virtual environment, as during hardware failure, the running virtual machines are started on another host machine and the downtime is reduced to minimal. If a server fails, affected virtual machines are re-started on other production servers that have spare capacity. In datacenter, this feature would give high availability to the virtual machines by starting them on other servers and thus minimizing the impact on failures.

Using the bare-metal hypervisor makes it simpler and less expensive to provide higher levels of availability for important applications. Using hypervisor, the servers in the infrastructure can easily increase the baseline level of availability provided for all applications, as well as provide higher levels of availability more easily and cost-effectively.

By implementing the High Availability (HA) feature for any datacenter, it is possible to reduce both planned and unplanned downtime. HA, a feature of bare-metal hypervisor, specifically reduces unplanned downtime by leveraging multiple hypervisor servers configured as a cluster to provide rapid recovery from outages as well as cost-effective high availability for applications running in virtual machines.

HA feature protects application availability against Hardware failure by restarting the virtual machines on other hosts within the cluster. Protection against operating system failure is obtained by continuously monitoring a virtual machine and resetting it in the event that an operating system (OS) failure is detected. Unlike other clustering solutions, HA provides the infrastructure to protect all workloads within the cluster: There is no need to install additional software within the application or virtual machine. HA protects all workloads that are in the infrastructure. After HA is configured, no actions are required to protect new virtual machines. They are automatically protected.

The following are the advantages when we configure the HA compared to traditional fail-over solutions:

- Minimal setup.
- Reduced complexity (e.g., no need for quorum disks).
- Reduced hardware cost and setup.
- Increased application availability without the expense of additional idle failover hosts or the complexity of maintaining identical hosts for failover pairs.

The datacenter can be supported with load balance feature as it is virtualized with bare-metal hypervisor. The action taken by HA for virtual machines running on a host when the host has lost its ability to communicate with other hosts over the management network and cannot ping the isolation addresses is called host isolation response. The word host isolation does not necessarily mean that the virtual machine network is down, but only that the management network, and possibly others, is down. If server monitoring is in disabled mode, restart of virtual machines in that server is also disabled on other hosts following a host failure or isolation. Essentially, a server will always perform the programmed server isolation response when it detects that it is isolated. The server monitoring setting determines whether virtual machines will be restarted in other servers in the same cluster following this event.

Fault Tolerance

Fault Tolerance feature of the virtualized datacenter leverages the well-known encapsulation properties of virtualization by building HA directly into the bare-metal hypervisor in order to deliver hardware style fault tolerance to virtual machines. This feature is to be used for all the virtual machines that require 100% uptime.

Dynamic Movement

Dynamic movement of virtual machines in the virtualized datacenter machines could give more options to do load balancing and hardware maintenance. Usage of this feature does not have any impact on the services offered by the virtual machine. This functionality is used by Distributed Resource Scheduling algorithm. Virtual dynamic motion enables the capability of live migration of running virtual machines from one physical server to another with zero down time, continuous service availability, and complete transaction integrity. Storage dynamic movement enables the migration of virtual machine files from one data store to another without service interruption. One can choose to place the virtual machine and all its disks in a single location, or select separate locations for the virtual machine configuration file and each virtual disk. The virtual machine remains on the same host during Storage Dynamic Movement.

This could be achieved using Distributed Power Management algorithm which will help to reduce energy consumption in the datacenter by optimizing workload placement for low power consumption with Distributed Power Management. It consolidates workloads when Distributed Resource clusters need fewer resources and powers off host servers to conserve energy. When resource requirements increase, Distributed Power Management algorithm brings hosts back online to ensure that service levels are met.

In any datacenter this feature would be effectively used, while provisioning the machines on demand. Distributed Power Management algorithm would also use this feature to save energy during the off peak hours. Migration with dynamic movement aids moving of a powered-on virtual machine to a new host. Migration with dynamic movement allows moving a virtual machine to a new host without any interruption in the availability of the virtual machine. Migration with dynamic movement cannot be used to move virtual machines from one datacenter to another.

Dynamic Storage

Dynamic movement of virtual machines along with the virtual hard disks in any datacenter machines could give more options to do load balancing and hardware maintenance of storage devices. This feature allows administrators to move the virtual disks or configuration file of a powered-on virtual machine to a new data store. Migration with storage dynamic movement allows moving a virtual machine's storage without any interruption in the availability of the virtual machine. Usage of this feature does not have any impact on the virtual machine.

Resource Scheduler

In the virtualized datacenter, the presence of a Resource Scheduler algorithm would improve resource allocation, efficiency, and power consumption in virtual infrastructures. Resource Scheduler balances workloads according to available resources, and users can configure Distributed Resource Scheduler algorithms for manual or automatic control. If a workload's needs decrease drastically, Distributed Resource Scheduler can temporarily power down unnecessary physical servers.

Resource Scheduler works with Virtual Dynamic Motion to provide automated resource optimization and virtual machine placement and migration, to help align available resources with pre-defined business priorities while maximizing hardware utilization. Distributed Resource Scheduling algorithm simplifies the job of handling new applications and adding new virtual machines, simplifies the task of extracting or removing hardware when it is no longer needed, or replacing older host machines with newer and larger capacity hardware. Adding new resources is also straight forward, as one can simply drag and drop new physical hosts into a cluster.

A Distributed Resource Scheduler cluster is a collection of physical bare-metal hypervisor installed servers and associated virtual machines with shared resources. When somebody adds a host to a resource scheduler cluster, the host's resources become part of the cluster's resources. In addition to this aggregation of resources, with a Distributed Resource Scheduler cluster can support cluster-wide resource pools and enforce cluster-level resource allocation policies allowing to dynamically provision compute resources to meet the demand in an efficient way while retaining the SLAs.

Distributed Resource Scheduler algorithm provides automatic initial virtual machine placement on any of the hosts in the cluster, and also makes automatic resource relocation and optimization decisions as hosts or virtual machines are added or removed from the cluster. Distributed Resource Scheduler algorithms can also be configured for manual control, in which case it only makes recommendations that can be reviewed and carried out.. The Distributed Resource Scheduler and Dynamic movement integration combination would make the infrastructure a redundant one and thus minimize the impact in an event of failure.

Power Management

Usage of a power management options in virtualized environment would significantly improve efficiency, thereby reducing the power consumption for virtual infrastructures. Power management application balance workloads according to available resources and users can configure this feature along with Resource Scheduler. If a workload's needs decrease drastically, scheduling algorithms can temporarily power down unnecessary physical servers using Distributed Power Management algorithms. These servers are brought back online automatically when there is a requirement for more compute resource.

Provisioning and De-Provisioning

Any datacenter infrastructure can be virtualized and the option of provisioning comes along with it for creating a virtual machine. The simplest reason for using virtual machine templates is efficiency. By using the templates, many repetitive installation and configuration tasks can be avoided. It is to be noted that a datacenter can utilize the capabilities of hypervisor and virtual management server for the automatic provisioning and de-provisioning functionality by making the infrastructure virtualized. The option of provisioning comes along with it for creating a virtual machine. The simplest reason for using virtual machine templates is efficiency. By using the templates, many repetitive installation and configuration tasks can be avoided. The outcome is a fully installed, ready to operate virtual machine in less time than that required for manual installation with all the features and configurations as the source machine. On-demand provisioning of the resources for de-duplication process and provisioning more servers require resources on demand basis. Also hypervisor should be able to scale up and down the infrastructure as per demand.

Moreover hypervisor should also be able to detect new hardware such as server, storage, etc. that are being introduced into the existing infrastructure. It should also maintain the balance of the resources in the cluster. HA of the machines that are hosted in the virtual infrastructure should also be guaranteed.

Dynamic Allocation and De-Allocation

Virtualized datacenter is scalable and capable of using the existing resources in an efficient way. This is achieved with the bare-metal hypervisor that is installed on the servers in the datacenter. This environment will not be only scalable but intelligent enough to understand the load on the datacenter and allocate the computing resources accordingly. This would save significant amount of energy and will also be able to use the existing computing resources in the datacenter effectively. During the off peak hours, similarly lots of computing power would be in unusable state. The Distributed Power Management algorithm with the help of Distributed Resource Scheduler algorithm would identify the less resource consumed servers. Using dynamic movement, the virtual machines running in that server would be moved dynamically to the other servers. Then the server is moved to power off state by communicating through the remote console. These servers are brought online as and when the requirement for the computing resources arises. This would save a significant amount of energy in terms of power and computing resource, this enabling a GO Green Datacenter.

The load balancing is being done efficiently for the peak hours and non-peak hours. The bare-metal hypervisor is the one that makes the available computing resources to be used effectively and efficiently. Templates can be a time-saving feature for virtualization administrators as they allow cloning, converting, and deploying virtual machines. A template is a 'golden' copy of a virtual machine (VM) organized by folders and managed with permissions. They're useful because they act as a protected version of a model VM which can be used to create new VMs. As a template is the original and perfect image of a particular VM, it cannot be powered on or run.

By using Distributed Power Management algorithm along with the Distributed Resource scheduler, the multiple datacenters could be optimized of power usage by moving the unused physical machines to standby mode. The challenge here could be to understand which Physical machine needs to the turned off. Resource Scheduler algorithm should have the ability to understand that free and used resource capacity in a cluster. Using algorithm it will move the VMs running on one physical host to another to make one physical host completely offline. This is done automatically and dynamically by Resource Scheduler algorithm, as there is no service loss to the end user. This feature also allows an administrator to define the rules and policies according to the priority which decides how each VM should share resources and how the available resources are prioritized among multiple virtual machines. It also sends the heartbeat signal to all the hosts to ensure that it is up and running fine. So this feature is capable of dynamically provision, resource quickly as and when needed when resources are free in the cluster.

7.4 SERVER VIRTUALIZATION

Server virtualization covers different types of virtualization such as client, storage, and network virtualization. In this section, different implementations of virtualization, management software, what constitutes support for virtualization platforms, and other related topics like appliance and cloud computing are discussed.

Server virtualization is the masking of server resources, including the number and identity of individual physical servers, processors, and operating systems, from server users. The server administrator uses a software application to divide one physical server into multiple isolated virtual environments. The virtual environments provide an abstraction of a complete, independent server to the server users (Figure 7.1).

7.4.1 Virtual Machine

This is often called virtualization environment, virtualized environment, partition, or container. A virtual machine (VM) is a server environment that does not physically exist but is created within another server. In this context, a VM is called a 'guest' while the environment

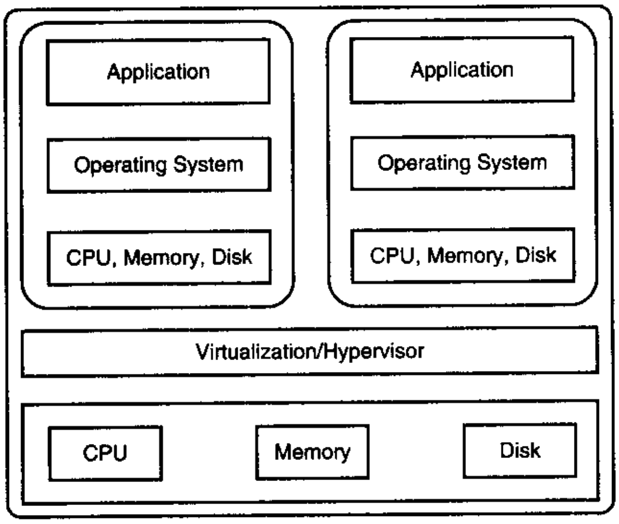

FIGURE 7.1 Server virtualization.

it runs within is called a 'host.' One host environment can usually run multiple VMs at once. Because VMs are separated from the physical resources they use, the host environment is often able to dynamically assign those resources among them.

A user interacting with a VM can view it as a physical machine, in the sense that the user would see access to an operating system and machine resources like CPU, memory, hard disk, and network. For instance, a hypervisor virtualizes a server with architecture into multiple virtual machines. Each VM is a virtualized server with its assigned system resources and an operating system.

7.4.2 Virtualization Technologies

Two major types of technology are employed in server virtualization: hardware virtualization and OS virtualization. Hardware virtualization virtualizes the server hardware, and OS virtualization virtualizes the application environment (for example, file systems).

7.4.3 Hardware Virtualization

Hardware virtualization is also known as Hypervisor-based Virtualization, Bare-metal Hypervisor, Type 1 Virtualization, or simply Hypervisor. This virtualization technology has a virtualization layer running immediately on the hardware, which divides the server machine into several virtual machines or partitions with a guest operating system running in each of the machines (Figure 7.2).

This virtualization approach provides binary transparency because the virtualization environment products themselves provide transparency to the operating systems, applications, and middleware that operate above them.

7.4.4 OS Virtualization

This type of server virtualization is also known as OS-based Virtualization, OS-level Virtualization, or Type 2 Virtualization. OS virtualization creates virtualization environments

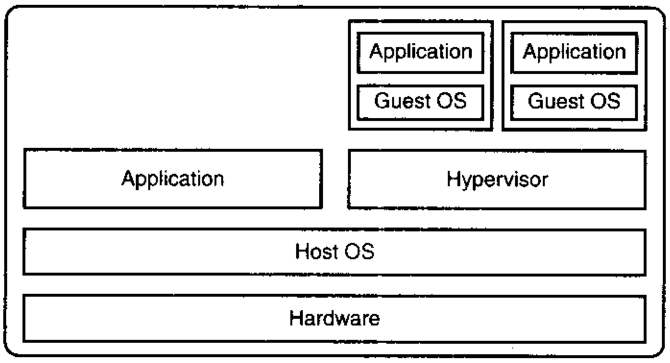

FIGURE 7.2 OS virtualization.

within a single instance of an operating system. The virtual environments created by OS virtualization are often called containers.

Because all virtualization environments must share resources of a single operating system while having a private virtual operating system environment, a particular implementation of the technology may alter file system orientation and often introduce access restrictions to global system configuration or settings.

7.5 VIRTUALIZATION FOR x86 ARCHITECTURE

Virtualization on processors encounters a set of challenges that the virtualization on RISC processors does not have. This is mainly because the vendors or technology providers for processors, systems, virtualization technologies, and operating systems are different and operate independently. As a result, the virtualization technologies and the rest of the system are available separately and on different timelines rather than as a single integrated unit. Therefore, both forward and backward compatibilities must be considered when designing virtualization for x86.

Most operating systems, including those for x86 such as Windows and Linux, are designed to run directly on the bare-metal hardware, so they naturally assume that they fully 'own' the computer hardware. The x86 architecture offers four levels of privilege known as Ring 0, 1, 2, and 3 to operating systems and applications to manage access to the computer hardware. While user-level applications typically run in Ring 3, the operating system needs to have direct access to the memory and hardware and must execute its privileged instructions in Ring 0.

Virtualizing the x86 architecture requires placing a virtualization layer under the operating system (which expects to be in the most privileged Ring 0) to create and manage the virtual machines that deliver shared resources.

Hardware-based virtual machine and paravirtualization are ways to overcome the challenges.

7.5.1 Paravirtualization

Also know as **OS-Assisted Virtualization**, this term is used to describe the virtualization techniques used to overcome the virtualization challenges on older versions of processors. The technique requires modifying the guest OS kernel to improve the communication and performance between that kernel and the virtualization layer hypervisor. The newer versions of processors provide on-chip virtualization features called hardware-based virtual machine that make paravirtualization unnecessary.

Paravirtualization involves modifying the OS kernel to replace non-virtualizable instructions with hypercalls that communicate directly with the hypervisor. Paravirtualization also allows a set of kernel operations to be bypassed in favour of a hypervisor call that encapsulates the entire set. As such, it adds value beyond simple instruction emulation.

7.6 HYPERVISOR MANAGEMENT SOFTWARE

For each hypervisor, there is a companion layer of hypervisor management software that provides a range of functions like create VM, delete VM, move VM, etc. as the hypervisor management function controlling the hypervisor. A unique set of APIs and GUIs is available for each 'Hypervisor/Hypervisor Management Software' pair that is used by the client IT staff and by ISVs to create management services or other applications.

7.6.1 Hypervisor

Hypervisor is the foundation for virtualization on server, enabling hardware to be divided into multiple logical partitions and ensuring isolation among them. This also supports Ethernet transport mechanism and Ethernet switch which are needed for VLAN capability. VLAN allows secure communication between logical partitions without using any physical Ethernet adapter. Hypervisor supports Virtual SCSI to provide support for virtual storage.

Hypervisor is a global firmware image located outside the partition memory in the first physical memory block at physical address zero. Hypervisor takes control as soon as the system is powered on and gathers information about memory, CPU, I/O, and other resources that are available to the system. Hypervisor owns and controls all the mentioned resources and other-resources that are GLOBAL to the system. Hypervisor performs virtual memory management using a global partition page table and manages any attempt by a partition to access outside its allocated limit. The whole physical memory is divided into blocks called physical memory blocks (PMBs). The logical memory is divided into logical memory blocks (LMBs). PMBs are mapped to LMBs. The Hypervisor has access to entire memory space and maintains memory allocation to partitions through a global partition page table. Service partition is a partition that is allowed to update the Hypervisor, which is a processor-based firmware. It is the nerve center of the Virtualization Engine. This handles micro-partitioning of the CPU and the memory pool.

7.7 VIRTUAL INFRASTRUCTURE REQUIREMENTS

Virtualization products have strict requirements on back-end infrastructure components, including Storage, Network, Backup, Systems Management, Security, and Time Sync.Ensuring that these existing components are of a supported configuration is critical to the success of the implementation. During this engagement, an IT Architect reviews and documents the current environment, and where applicable, make recommendations on changes required to optimize the infrastructure.

Where applicable, enterprise tools are used to gain a clear understanding of the environment and the configuration and utilization of various systems. A virtualization sizing tool is then used to accurately calculate the size of a potential virtualization platform.

7.7.1 Server Virtualization Suitability Assessment

One of the key advantages of virtualization is greater utilization of physical server resources. Achieving this advantage must not be at the cost of service to the business. It is vital to ensure that the virtualization host server is sized such that it can deliver acceptable levels of service to all guests.

To ensure that existing servers operate in a shared environment, detailed hardware inventory and performance utilization information must be obtained, and extrapolated and analyzed for suitability and host server sizing.

At the completion of the collection phase, the architect evaluates the results and provides documented recommendations on virtualization suitability across the server candidates.

7.7.2 Detailed Design

Virtualization introduces many changes into the environment, and ensuring that the platform co-exists and interacts with the existing infrastructure is the key to a successful implementation.

The purpose of the design is to set naming and security standards, define the disk and network structure, document any required system tuning elements, and produce a virtual infrastructure design capable of meeting your specific requirements for a virtualized Intel server environment.

Detailed Design Document

Virtualization design document should include the following:

- Security and administration model.
- Backup methodology.
- Host physical and virtual disk layout, specifically around file system structure, and-dedication of disks to guests where applicable.
- Virtual network topology structure/format and inter-connection with the physical network.
- Virtualization service console configuration.
- Virtualization kernel device share factor configuration.
- Host server hardware specifications.
- Virtualization management server configuration, including database and directory services integration.
- Virtual machine distribution amongst hosts.
- Processes and procedures for ongoing management.
- Implementation tables and configuration settings.

7.8 SUMMARY

This chapter focuses on server virtualization but also covers other types of virtualization. Under the server virtualization, we have discussed different implementations of virtualization, management software, what constitutes support for virtualization platforms, and other related topics.

CHAPTER 8

CLOUD INFRASTRUCTURE: DEEP DIVE

Introduction

Storage Virtualization

Storage Area Networks

Network-Attached Storage

Cloud Server Virtualization

Networking Essential to Cloud

Summary

8.1 INTRODUCTION

Businesses continually seek ways to reduce cost and risk while increasing the quality and agility of their IT infrastructure. Especially for server hardware, they are always looking for new ways to help improve overall utilization and to increase the flexibility with which they can deploy their hardware to meet the ever-changing business needs.

Virtualized IT environments and cloud computing will put new requirements on networks, both inside and to and from datacenters. Networks will require new levels of performance, availability, resiliency, security, and management while delivering the cost-effectiveness and energy efficiencies expected from the rest of the infrastructure. Without the right networking infrastructures, businesses will not be able to realize the benefits of virtualization fully. The new services are designed to help networks adapt to the new demands of virtualized infrastructure and condition the IT infrastructure for cloud computing.

What is happening in the datacenter today is the adoption of virtualization. It helps businesses get better utilization out of the resources they have, makes them easier to manage, and saves money.

However, virtualization and its benefits are being adopted in multiple areas of the IT infrastructure from the different server platforms, x86, RISC, and mainframes and the different hypervisiors, to storage and even the network. Businesses need help in understanding how the different virtualization techniques will affect their network and how they should plan and design their future network.

Virtual machines enable businesses to run multiple operating systems concurrently on a single physical server, providing for much more effective utilization of the underlying hardware. There are various scenarios like software testing and development, legacy application re-hosting, server consolidation, and testing of distributed server applications on a single server when developer and server administrator can reap values from server hardware solutions.

The tremendous growth in data over the last decade needs lot of control mechanisms. But the IT business rule that works for IT environment in which computing is decentralized have not controlled the storage, processor, and networking requirements. This has made the system very complex and the data available is fragmented over the legacy systems. This needs a complete lifecycle management for cloud environment to help the cloud subscribers. This will make the storage, archival, and information dissemination easier for the business operations.

Let us start with the facts:

- Data is growing rapidly, approximately by 50 percent every year.
- The companies are running big amount of storage and some industries like health and life sciences needs even ITB data per day.
- The total IT budgets comprise around 15 percent because of storage.
- The redundancies are higher.

There are a number of problems at the enterprise-level: Now companies are not able to control the rapid growth of the data resulting in a lack of efficient storage and information management systems. Therefore, it adds more costs and gives rise to the problems of not meeting the service level agreements. These problems lead to insignificant performance over legacy systems.

Indeed, there is requirement of the information lifecycle management techniques to overcome these problems and meet the future high volume data growth problems.

As company data loads continue to increase, so do the complexity and capacity of the storage environment. Storage area networks (SANs) can help overcome some of the storage challenges, but not all of them. The complexity remains. Multiple storage devices from multiple vendors can have interoperability issues on the SAN. Storage management can be tedious and time consuming. And infrastructure changes can be difficult to implement.

By embracing storage virtualization, clients gain the ability to reduce storage network complexity by aggregating multiple storage devices into a common, managed virtual storage pool. And Storage Optimization and Integration Services – storage virtualization service products – help businesses create an integrated, virtualized solution that aligns with their unique storage strategy, vendor choices, and environment. By gaining access to experienced, knowledgeable professionals who use proven methodologies, businesses can develop a storage virtualization solution that allows them to add the storage solutions with live up-gradation means without effecting the actual servers and network. This will even help to develop the comprehensive approach to combine with varied speed based drives based on size requirements may be from different vendors. It provides ease of access of storage devices across the enterprise.

8.1.1 Value Proposition

For organizations that need to reduce storage management complexity, increase storage capacity utilization, and enable non-disruptive hardware change, cloud vendors offer storage virtualization services.

Designed to help clients reduce storage costs, centralize data management, and extend the useful life of their storage hardware, this service product provides experienced consultants to design and build an integrated virtualization solution that aligns with clients' particular storage strategies and environments.

Unlike many of its competitors, cloud vendors are multi-vendor suppliers and integrators that use a comprehensive consulting approach and can provide the expertise, techniques and broad product portfolio their clients need to create an efficient virtualized storage environment.

8.2 STORAGE VIRTUALIZATION

Storage virtualization improves the utilization of storage and people assets because it allows you to treat resources as a single pool, accessing and managing those resources across your organization more efficiently, by effect and need rather than physical location (Figure 8.1).

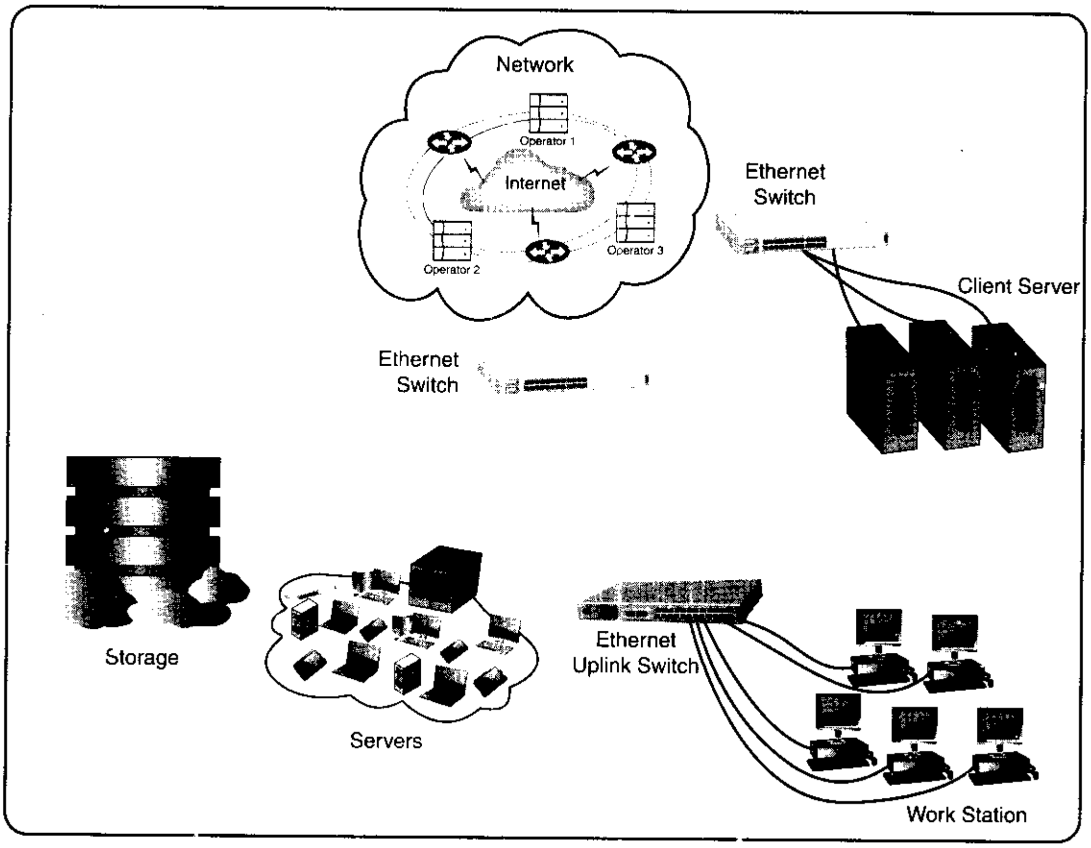

FIGURE 8.1 Storage cloud.

Benefits

- Make storage simpler.
- Make storage more heterogeneous.
- Make storage more manageable.

Why Cloud?

- Cloud can assess, design, develop, optimize, and support on-demand infrastructures that are integrated, virtualized, and autonomic and are built on open standards.

The Storage Challenge: Storage is a top priority for every business – mission-critical as well as challenging to manage. What makes it challenging?

- Growth in storage demand and therefore growth in storage management costs due to digital content, e-mail, Internet-based applications, and emerging technology.
- Threats to business continuity posed by disaster – even human error – cause of 40 percent of outages. Dealing with storage management strains your budget.

- Pressure to retain data for compliance with regulations has increased worldwide.
- Complexity of storage networks with devices from different manufacturers resulting in separate islands of storage is on the rise.

A single point of management over your entire stoiage netwo:k, using your storage and data resources to their full potential, among others, enables excellent productivity of storage administrators and increases the potential to reduce errors. Typical structural client savings of 30 to 70 percent on storage management costs are possible.

8.2.1 Storage Cost Drivers

Storage is Growing Rapidly: Although cost of storage hardware is decreasing (halving every 12 months), the overall storage cost is increasing as a result of increased demand for storage (nearly doubling every 12 months) and complex storage management:

- Only 14 percent of the total cost of ownership (TCO) is hardware cost.
- Studies indicate that storage-related cost (hardware and software) will peak to 23 percent of the IT budget.
- Today storage administrations are islands of point solutions, which increases management cost.
- While the purchasing cost per GB goes down by 20–40 percent yearly, the cost of managing the storage may rise high with growth in traditional storage environments.

8.3 STORAGE AREA NETWORKS

Storage Area Network (SAN) is a method of provisioning by locally attaching the device to the operating system, to the servers. With the SAN architecture, we can connect different types of disk arrays, tapes, and other storage devices (Figure 8.2).

Network-attached storage (NAS) is different with respect to SAN as it uses the file-based protocols such as NFS. In this architecture, it is evident that the storage is available remotely and can be accessed as a file and not as the disk block.

SANs are becoming a pervasive technology. This text shows the evolution through several stages:

1 – Direct-attach Storage:

Hosts can only use storage that is directly attached through point-to-point SCSI connections. The disk storage is physically separate and cannot be configured to attach to multiple hosts.

2 – Centralized But Still Direct-attach Storage:

The storage is physically centralized in one unit. The disk controller is connected to multiple hosts with point-to-point connections. Re-assigning storage to different hosts requires re-cabling.

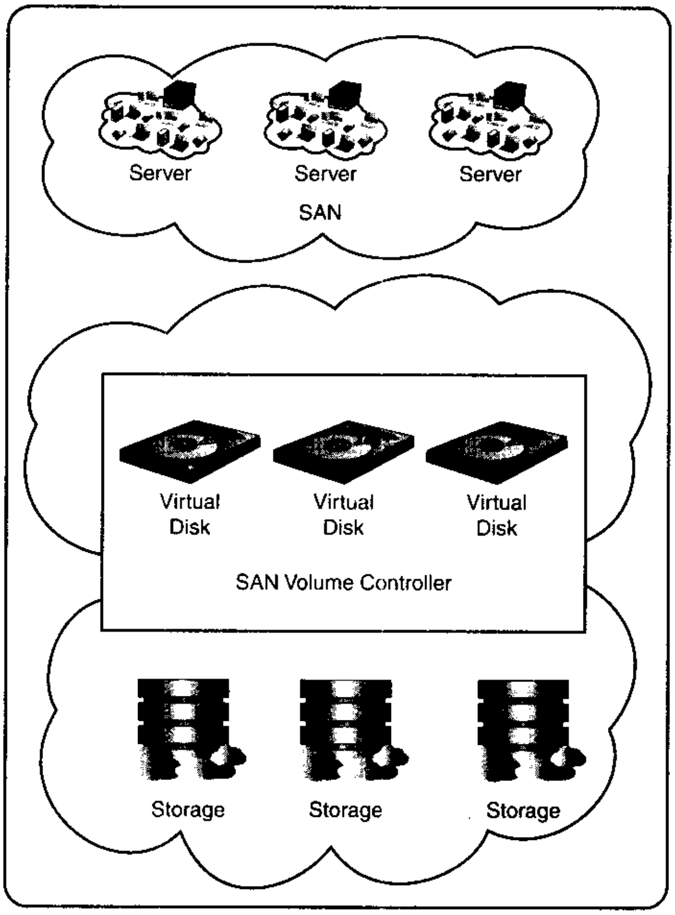

FIGURE 8.2 SAN.

3 – Shared Storage:

A SAN enables point-to-point connections between any host and disk controller. Disk volumes can be dynamically assigned to different hosts, without requiring re-cabling. SANs have enabled several benefits, including:

- Better connectivity (avoidance of re-cabling).
- Improved performance (through higher bandwidth connections).
- Distance flexibility (overcoming SCSI limitations).
- Scalability (enabling storage capacity to be increased with less disruption).
- More vendor/product choice (by separating the disk storage from the host server and using open standard Fibre Channel connections).

However, these benefits have also resulted in new issues and increased complexity. Let's see what this means:

Complexity Case A: Enterprises have a SAN with many UNIX/Windows servers attached.

Complexity Case B: Within the SAN are many different types of storage – in some cases, customers made this choice to stay vendor-neutral; in other cases it was a result of mergers or consolidations.

The storage administrator has to configure LUNs to servers and keep track of which servershave what storage. Surprisingly, most customers admit that they are keeping track of all this with spreadsheets as SAN managers are not as prevalent as we had thought. You can imagine the complexity of reallocating LUNs to different servers as the need for storage shifts from one server to another.

Complexity Case C: Complexity of different file systems. Now the storage administrator needs to know the different commands to use depending on the file system being dealt with.

Complexity Case D: On top of all of this, each of these storage devices has to be configured and installed. But because there are no common standards, each storage device has its own procedures and user interfaces for doing this. Now imagine that you're in production and the storage administrator is faced with the task of doing replication of all of these devices across all these servers and managing the performance and capacity of each device. Again, they each have their separate interfaces and procedures to do this.

Complexity Case E: Last but not the least, since the file systems are really tied to each of the hundreds of servers, all of the storage management functions have to be run on hundreds of servers. If you ask businesses what percentage of their storage is actually being utilized, most of them will not know the answer.

So, this is what businesses mean when they say they have inter operability and manageability problems with their storage. No wonder, if you think about the many permutations and combinations of "x" servers/operating systems times "y" devices types, you know that the complexity is daunting.

8.3.1 Storage Virtualization Benefits

Storage virtualization makes storage simpler, more heterogeneous and more manageable. It creates a logical view of storage that simplifies management and makes physical changes to the infrastructure transparent to the user (Figure 8.3).

- **Make It Simpler**
  - Effective use of capacity.
  - Effective management of capacity.
  - Lower total cost of ownership.
- **Make It More Heterogeneous**
  - Any to any attachment.
  - Data migration.
  - Security.
  - Investment protection.
- **Make It More Manageable**
  - Scalable.
  - Quality of service management.

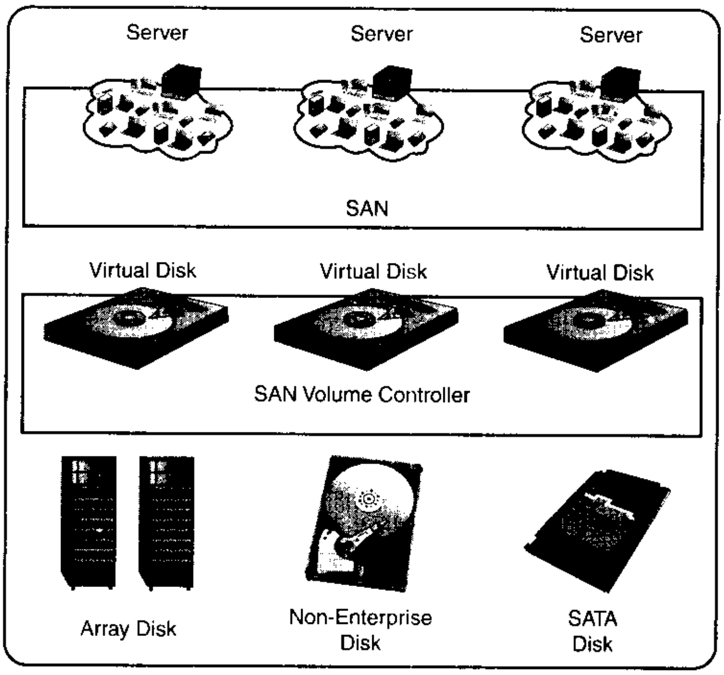

FIGURE 8.3 Virtual storage infrastructure.

This gives the options to control the storage volumes through the central point. It even reduces the server downtime for different predicted outages, maintenance, and support activities. This helps resource utilization and sets the storage management tools in a cost-effective fashion.

8.4 NETWORK-ATTACHED STORAGE

Network-Attached storage (NAS), based on its network address, is a hard drive storage. It is not directly connected to the server that actually serves the workstations connected to the network applications.

It is advantageous to separate the storage with the server as it makes the process faster because both the application and files are not challenging the same resources on the network. NAS works on local area network based on Ethernet switch with the IP address. There exists the mapping between the main server and file server.

NAS comprises Redundant Array of Independent Disks (RAID) systems, hard disk storage and configuration, and file mapping with network-attached device management device. NAS works on file-based protocols. This can comprise NFS, SMP, CIFS, etc.

8.4.1 NAS Basics

NAS is different from the file server based on the operating system. These file servers are based on the offerings of the file servers like UNIX, Novell, Linux, Windows, and OS/2. NAS does not have the feature of application or directory server but it is based on the specific function (Figure 8.4).

NAS uses a plug-and-play feature to the Ethernet network and makes it available within the fraction of seconds. NAS delivers the fully loaded performance-based application environment to serve the files available on the network systems. These systems can be connected to the non-Windows based environments to attach to the files system like NFS and serve the large ports.

Implementation of NAS architecture is not difficult but administrator should know the gamut of the components of the NAS like what are the interconnection points and NAS protocols.

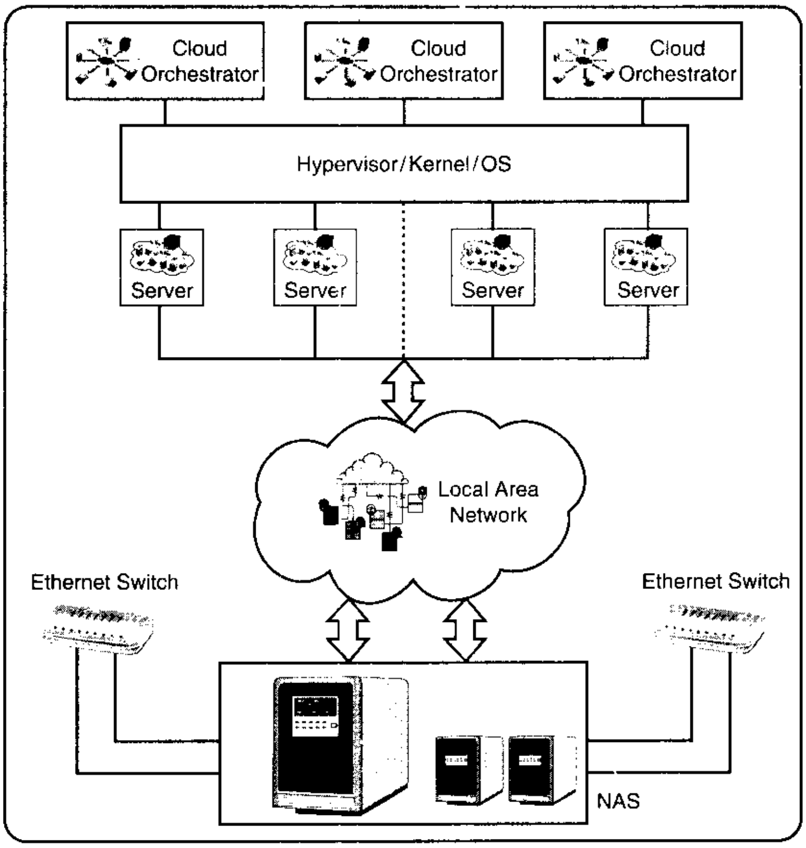

FIGURE 8.4 NAS.

8.4.2 NAS Protocols

We need a language to talk to the NAS interconnects. It is important to have a knowledge of the protocol that is the key for NAS implementation.

Common Internet File System started as the project outcome of Server Messaging Block Protocol and was later renamed as Common Internet File System. Common Internet File System is the most common protocol for windows environment.

Another protocol for the NAS is Network File System (NFS). The combined power of virtualization and performance gives the upper hand to use NFS as a preferred protocol among NAS vendors. This helps in storing the data at the file level rather than at the block level.

iSCSI is an another protocol stack for NAS and gives the options to access the data at block level. This is an inexpensive protocol compared to the other two as it works without using expensive adapters and sophisticated software.

Fibre Channel over Ethernet (FCoE) is another NAS protocol. It is a combination of Fibre Channel and Ethernet but it is difficult to tell about its success as there are not that many Fibre Channel over Ethernet deployments.

8.4.3 NAS Interconnects

In order to integrate NAS devices, we need NAS interconnects. The more we understand the NAS interconnects, the easier it will be to administrate for better storage decisions about NAS systems.

Fast Ethernet is one of the best among all the interconnect architecture. It is very easy to understand and not expensive. It is important to note that it works slowly at basic file levels. It is good for transfer of up to 10 MB of data with a group of fewer people. The enhanced and more advance level of interconnect after Fast Ethernet is Gigabit Ethernet.

Gigabit Ethernet serves to a larger group of people with the file transfer of more that 10 MB. This can accommodate 100 people accessing the server at a very good speed and relatively very good performance.

Gigabit Ethernet is a good interconnect protocol, but when we increase our group size, to say 500, we can find some performance issues. It may also have issues for block-level accesses.

In order to overcome these type of problems, the relatively new interconnect protocol, 10 GE, is of help. This is the fastest interconnect protocol currently and provides very good performance when there is demand of huge size file transfers with more speed. Enterprises use hybrid interconnect protocols as per the requirements. Price plays a vital role while choosing the interconnect protocol.

8.4.4 NAS Requirements

NAS requirements need some basic steps to get the best out of a situation. It will help you to define the success for the NAS implementation. It will help to maintain the most suitable set of vendors based on the requirements of the NAS systems.

It is a good idea to understand NAS implementation before starting it. All the vendors have different devices available over different deployments so having an idea for the same gives an upper edge for the best fit of the NAS system. It will be good to evaluate the different options available by distributing the requirements among the vendors and outlining the different extremities of the situation.

Listing of the requirements required for the device can be the easiest step. The primary requirements can be connecting to the multiple servers without any performance loss, the amount of the storage to handle the file use, etc. This will require little bit of research on what are the current offerings available to handle the requirements. The requirements that are technical necessity of the NAS systems can be maintained as a list to match up with the real-world deployment and budget management to buy it.

8.4.5 High-Performance NAS

NAS works with file-based systems that help companies to work with distributed file systems with the help of file systems like NFS and CIFS. This consolidates the distributed file system into different, small file-based storage systems. There were some problems associated with NAS like reliability, connectivity, and scalability with the enterprise storage. But now the new generation NAS systems have overcome it and promise higher value in the same file-based storage systems.

In order to differentiate the high-performance from regular NAS, there is no common single point of agreement. But we can have some distinction at least such as high-performance NAS provides more ports on the interface level. Connectivity is an important factor for performance and more the number of ports, more reliable the NAS implementation.

High-performance NAS systems operate with the SATA or serial-attached SCSI. As a result, we get higher scalability with the disk controller engines available within the system. High-performance NAS systems work with various heads that can access various disks, and at the same time, improve performance. It is also possible to analyze the input/output operations per second (IOPS) to optimize the high-performance NAS system. We can have the concurrent file access in the high-performance NAS platform or access to multiple metadata points at the same time.

Clustering also plays an important role in increasing the throughput of NAS systems other than having the single pool of storage. It also offers resiliency as even if one of the cluster is not working, the other will not be affected and the workload of the failed one can be transferred to the working and free cluster.

The unique feature of high-performance NAS is Global File System (GFS). It is very helpful in the clustered environment where all the clusters share a single pool. GFS can be added with the operating system or can be added as a separate layer on top of the high-performance NAS. GFS helps cluster to work as independent or as a same entity while sharing the same pool.

High-performance NAS has an upper and helping edge over the deployment challenges. It helps:

- Perform more work in less time.
- Reduce the number of file servers.

- Simplify NAS storage infrastructure.
- Save energy.

Administrators should evaluate the prospective high-performance NAS system for:

- Underlying NAS systems management requirements.
- Expectation of NAS systems management efficiency.
- High performance and throughput issues.
- Demand of special host software or drivers.
- Expense of added software maintenance issues.
- Tuning and load balancing.
- Appropriate system for data workload.
- Management and interoperation with storage vendors to ease compatibility concerns.

The most important step for better performance results is to optimize the workload and understand the data workload with the inner depth of application. Like we can know what type of application it is and how it works – sequentially, or storing transactional data with the known IOPS.

High-performance NAS systems require the following to accommodate high-performance systems:

- New switching.
- Network architecture changes.
- Additional LAN bandwidth.

High-performance NAS is evolving fast but still many features of traditional NAS are not available in it yet, such as:

- Snapshots.
- Replication.
- Point-in-time (PIT) copies.
- Finer management granularity.
- Better load balancing and data migration.

High-performance NAS has also renewed interest in:

- Storage virtualization.
- Virtual machines and used throughout.
- Storage virtualization aggregation.
- Live storage mobility.
- High-capacity storage systems.

8.4.6 Network Infrastructure

Network infrastructure helps organizations to understand their networks and their network usage better, address specific networking issues or problems, and reduce networking costs.

The service provides relevant recommendations based on the analysis of data from the client's network and networking environment, industry insights, and leading practices in networking.

The service is composed of three activities from which the client may choose one or more activities on which to engage:

- Network infrastructure assessment.
- Network performance analysis.
- Network diagnostic assessment.

Network infrastructure assessment focuses on helping clients to understand the components that are deployed in their networks and includes reviews of their network designs, devices, and service level agreements and readiness reviews for deployments or refreshes with new technologies.

Network performance analysis and capacity planning focuses on helping businesses to understand how the performance of their network is being affected by business critical applications and background traffic, and to determine or predict the network sizing requirements to optimize the usage by applications.

Network diagnostic assessment focuses on helping businesses with problem determination, problem source identification, and root cause analysis of network performance problems.

Network infrastructure services can be appropriate any time your network is experiencing performance problems, you are under pressure to do more with fewer resources, or you have other networking issues to address. These services are also beneficial when you are planning to deploy important new business applications, bring in significant new traffic driven by organic growth or a merger or acquisition, introduce a new and network-dependent business model, or deploy a new technology such as voice or video over IP, wireless communications, or radio frequency identification (RFID).

Network infrastructure assessments, network performance and capacity planning studies, and root cause analysis of infrastructure performance problems help to prepare for new initiatives or to repair the current networking environment.

8.5 CLOUD SERVER VIRTUALIZATION

This section provides an overview of the architecture, features, and benefits of server virtualization solutions that provide a cost-effective virtual machine solution for OS platforms. Virtual machine technology enables customers to run multiple operating systems concurrently on a single physical server. Virtual Infrastructure Server Solution addresses a set of key customer scenarios, including consolidating and automating software testing and development environments, migrating legacy applications, consolidating multiple server workloads, and testing distributed server applications on a single physical server (Figure 8.5).

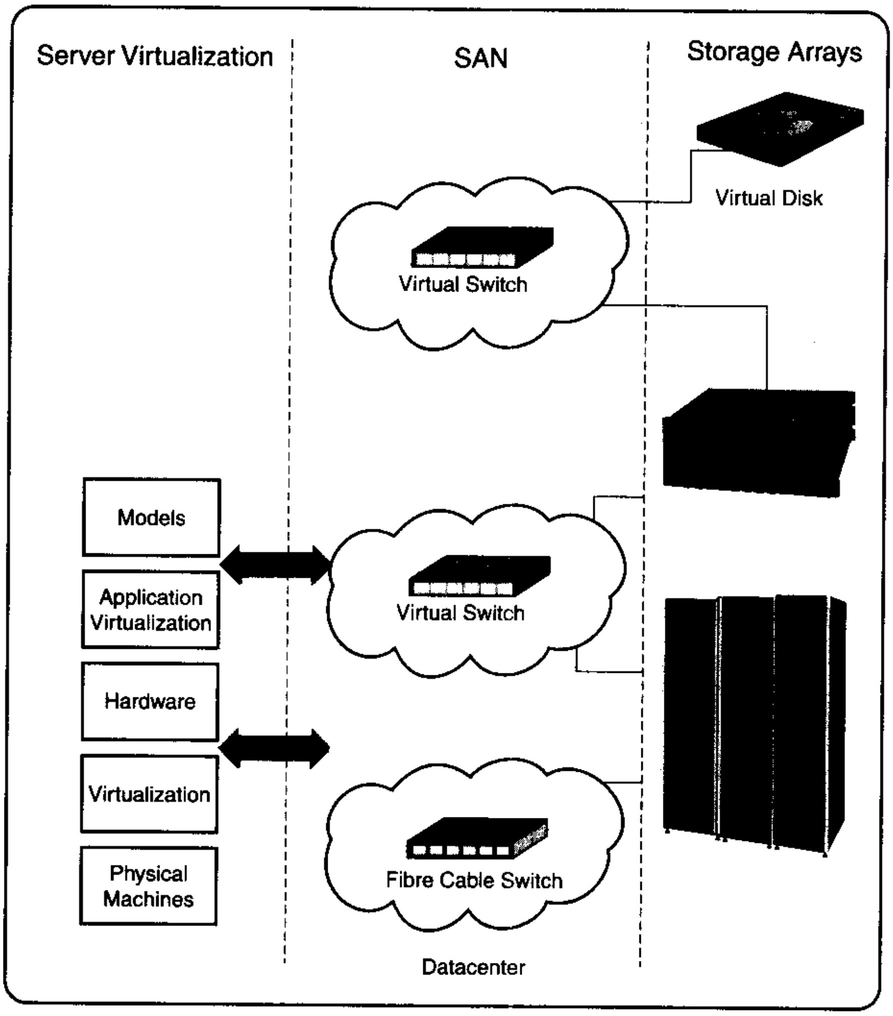

FIGURE 8.5 Server virtualization.

8.5.1 Datacenter Virtualization

The primary requirements of the datacenter with respect to datacenter virtualization are:

- Virtual servers.
- Storage.
- Networking.
- Unmodified operating systems.

These servers help to run independently the application on the virtual machines by sharing same pool of resources. They enable:

- Comprehensive virtualization.
- Management.

- Resource optimization.
- Application availability.
- Operational automation capabilities.

8.5.2 Virtual Datacenter

Organizations are always in need of higher level of utilization and flexibility of hardware resources. Virtual servers help achieve this by abstracting the memory, processor, storage, etc. available in the form of virtual resources.

The most important feature of virtual servers is that it provides:

- High level of performance.
- Scalability.
- Flexibility.

These virtual servers work like a complete server and fulfill the requirements of processors. The advance virtualization techniques ensure availability of the resources when there is the acute need of resources.

8.5.3 Virtual Datacenter Management and Control

Virtual datacenters provide all management and control functions of the environment under the same umbrella. These functions offer the following benefits:

- It uses a very simple provisioning method to allocate the virtual servers with the easy-to-use interface and templates to deploy the virtual server.
- Businesses need special attention at different intervals; virtual servers help to automate the operational needs of the deployment and set the alerts when they are required at most.
- It helps to schedule and alert the different management, control, and support functions with scheduled automated tasks.
- It is a very good tool for metering the utilization of processor memory and IOPS requirements with detail reports.
- It helps to set the customized roles based on the type of work as well as access permission to the resources using different tiers.

8.5.4 Dynamic Resource

In the virtual datacenter, resource requirements are not same every time. They have different spikes to meet the dynamic nature of the environment. Dynamic resources adhere to the requirements and allocate the computing resources dynamically to meet the business goals. This helps in:

- Monitoring utilization.
- Common resource pool maintenance.
- Matching the business needs and changing priorities.
- Making additional capacity available.

- Migrating live virtual machines.
- Dynamically allocating IT resources.
- Creating rules and policies to prioritize resource allocation.
- Granting IT autonomy to business organizations.
- Providing dedicated IT infrastructure to business unit by higher utilization achievement.
- Having centralized IT control over hardware resources.

8.5.5 High Availability

A virtual datacenter requires the features of high availability to provide cost-effective failover options. It should not be based on related operating systems and virtualization technologies. The following features for high availability make datacenter more robust and reliable:

- Failover options to protect applications.
- Consistent defense mechanism for IT infrastructure.
- Live workload transfer in case of failover.
- Alerting the administrator for stringent situation.
- Zero downtime to meet SLAs.

8.5.6 Live Migration

Similar to the high availability option we discussed the previous section, live migration helps the environment to run and gives unparalleled availability and flexibility to meet the requirements of the business goals.

It monitors the utilization of servers, storage, and networking and leverages the virtualization technology to move the virtual machine from one server to another at unprecedented situations. This is maintained by the set of files managed by shared storage-based file systems to access the virtual machine at different periods of intervals or simultaneously. Even the network is virtualized, which ensures the smooth migration process.

The main features/benefits it has are that it:

- Balances the workloads by transferring the under performing servers.
- Ensures live migration within no time and not even traced by the end-user.
- Maintains, manages, and supports the resource pools automatically.
- Provides ease of hardware maintenance with failover alerts and scheduled maintenance activities.

8.6 NETWORKING ESSENTIAL TO CLOUD

Network plays a vital role in infrastructure management to:

- Reduce costs.
- Improve service.
- Manage risk.

It is important to focus on infrastructure initiatives essential for reaping benefits like:

- Server.
- Storage hardware optimization.
- Technology enhancements.
- Service management improvement.
- Security.
- Resiliency.
- Optimizing the network (hardware, software, management).

Highly virtualized infrastructure-based clouds meet with demanding network requirements that can restrict the growth of the infrastructure management activities. There is a need of following features:

- More stringent network performance.
- Fast reliable access to virtualized resources.
- Flexible and adaptable networks.
- Application workload mobility.
- Response to variable capacity requirements.
- Security to support multi-tenancy.

8.6.1 *Datacenter Network*

Networking is essential to datacenter consolidation and virtualization initiatives that prepare for dynamic infrastructures and cloud computing. As organizations drive to transform their infrastructures to reduce costs, improve services, and manage risk, networking is an element that is pivotal to success. While many organizations continue to focus on server and storage hardware virtualization and provisioning, optimizing the network to support these initiatives is essential to ensure that maximum benefit is derived.

To Support These Networking Requirements, Businesses Will Need:

- Expertise to assess, plan, design, and implement networks with holistic consideration of servers, storage, application performance, and manageability.
- Different options of cost and different ranges of performance to match their needs.
- Technological expertise to design and deploy the security policies.
- Simple operational software to lower the cost and to integrate the network and manage it.

8.6.2 *Market Opportunity*

- **Cost control, high availability and performance, and robust security are business imperatives:**
  - Businesses are looking for cost containment and reduction in the datacenter while addressing challenges in responding quickly to rapidly changing business requirements

- **Datacenter networking technologies are changing fast and in-house staff does not always have adequate time and experience to take appropriate actions:**
  - Businesses need help, especially in new areas like infrastructure virtualization, private optical networking, and converged data and storage networking
- **Datacenter consolidations as well as server and storage virtualization impact the network in terms of new requirements for flexibility, performance, and security:**
  - A near-capacity and often difficult-to-manage datacenter networking infrastructure resulting from years of ad hoc changes and updates that can impede plans for server and storage virtualization

First, especially in today's economic environment, cost control and cost savings are key. Cloud vendors can help businesses save money by optimizing their current network through consolidation and virtualization of network devices, while helping them plan for the future. With the rapid changes occurring in IT infrastructure technologies, many businesses do not have the time or experience to understand how the network will be impacted. Most of the focus has been on consolidating and virtualizing servers and storage without thinking long term about the supporting network.

8.6.3 Datacenter Network Services

Datacenter network services help organizations to design and deploy solutions that:

- Prepare the datacenter network infrastructure to support important initiatives such as server and network consolidation, virtualization, and energy savings.
- Integrate other servers, storage systems, and existing networking infrastructure.
- Help save on energy, space, and management costs by moving to fewer networking devices that are better utilized.
- Prepare for new technologies while maintaining security and resiliency.
- Provides greater freedom to focus internal resources on critical operational concerns.
- Help reduce datacenter sprawl.
- Help to build differentiating advantage through improved efficiency and business innovation.

8.6.4 Data and Storage Network Convergence

The convergence of the data network and storage network into a single physical infra-structure will be attractive to businesses that want to lower costs and complexity in their datacenters without forfeiting high availability and performance. Data and storage network convergence eliminates duplicate infrastructure, reducing the required hardware components – adapters, cables, and switches – and resulting in savings on hardware expenditures, power, cooling, and space. At the same time, with a single physical infrastructure, deployment, upgrades, and management will be simplified, contributing to lower total cost of ownership.

The services include:

- Developing the network design.
- Selecting vendors and preparing a detailed design.
- Creating a roadmap for migration.

- Carrying procurement, logistics site preparation.
- Configuring, installing, and testing the network.
- Providing on-going maintenance support.

Comprehensive network infrastructure solutions designed to meet the changing requirements driven by consolidation and virtualization in support of dynamic infrastructures and cloud computing.

The service product consists of the following components:

- **Consolidation and Virtualization:** These help to consolidate the datacenter by designing and deploying the virtualized IT environments. Convergence of data and storage networking helps to design and deploy a converged data and storage infrastructure that allows quick achievement of the financial benefits of Fibre Channel over Ethernet Technology.
- **Private Networking:** This helps in establishing the private network and design and deploy the connectivity between the resources of different datacenters.
- **Security/Firewalls:** This helps to design and deploy network firewall technology to support the security requirements in today's datacenter network.

8.6.5 Network Infrastructure

Network infrastructure engagements are assessments that look at the performance, availability, resilience, and cost of the network. A major target audience for these services are clients who are facing pressure to do more with less or who are or experiencing network performance or availability problems. Another important target audience are clients who are planning changes such as the deployment of a new application that may require network redesign, anticipation of increased network traffic from organic growth or acquisitions or new network traffic flows driven by datacenter consolidation, green datacenter migrations, and server relocations.

Network infrastructure optimization includes four components: network infrastructure assessment, network performance analysis and capacity planning, network infrastructure for consolidation and virtualization, and network diagnostic assessment.

Network infrastructure assessments are designed to help organizations provide an in-depth view of an existing network infrastructure and identify gaps between the client's current and desired capabilities. These assessments are designed to determine if there are resources that are underutilized or untapped and if more value can be extracted from the investments a client has already made. They can also include reviews of network designs, devices, configurations, and service level agreements (including service level objectives) as well as an assessment of the readiness to deploy a new technology or upgrade the infrastructure.

Network performance analysis and capacity planning helps clients understand how overall network performance is affected by applications and background traffic, whether current or new. It also demonstrates how the network may react to failures or changes in configurations. It captures the traffic flows across the infrastructure to establish baselines and trends including usage patterns to provide inputs for simulation modelling or virtual testing of proposed changes in the infrastructure. Outputs of network performance and capacity planning can be used as inputs to make immediate changes to improve performance and capacity as well as determine network sizing requirements for a planned environment change or for setting appropriate service level agreements with carriers.

Network infrastructure for consolidation and virtualization is a service component that helps clients address the increasingly complex networking aspects of a virtualized IT environment by identifying cost savings and delivering an optimized networking infrastructure. The service can also help to plan, design, and build a networking infrastructure that contributes fully to a dynamic infrastructure – one that adapts quickly and effectively to business opportunities and rapidly changing demands.

A network diagnostic assessment helps clients identify problems, pinpoint sources, and analyze root causes of specific performance issues. This type of assessment can uncover problems such as traffic congestion, latency, suboptimal configurations, or the impact of application behaviour on the performance of the network. For converged networking environments, network diagnostic can determine the issues around VoIP and video real-time protocol issues so additional drill down came be identified. The objective is to recommend approaches and corrective actions that help the client maintain business-critical application performance at levels that support their business goals.

All engagements result in actionable recommendations for optimizing the networking infrastructure.

Pain Points:

- Rising costs and challenges in responding quickly to rapidly changing business opportunities.
- Critical business processes jeopardized by downtime, security breaches, or the poor performance of the networking infrastructure.
- Determining the networking alternatives that address immediate challenges and protect future flexibility.

Network Infrastructure Provides the Following Benefits:

- Identifying areas to cut costs through consolidation and virtualization.
- Planning a network that fully contributes to responsive IT environment.
- Removing network as bottleneck to meeting availability, security, and performance requirements.
- Leveraging expertise to plan and justify a dynamic networking infrastructure tied directly to business needs.
- Achieving the optimal balance of business needs, network enhancements, and cost savings using a proven, structured, and robust approach.

Business Impact:

- Increasing costs for a proliferation of hardware.
- Business constrained by IT infrastructure.
- Lost revenue due to customer dissatisfaction and reduced employee productivity.
- Poor business image, fines, and investigations due to security breaches.

- Risk of inexperienced staff making poor long-term, strategic decisions.
- Inability to achieve the right balance of business, cost, and network benefits.

The technique provides a structured approach for deploying business applications integrated with IT capabilities

8.6.6 *Datacenter Networking Services Enhancements*

There are three service products under Networking Strategy and Optimization Services:

- Enterprise Network and Communications Strategy and Planning.
- Network Application Optimization.
- Network Infrastructure.

There are four service components available under Network Integration Services datacenter networks:

- Consolidation and virtualization.
- Data and storage network convergence.
- Private optical networking.
- Security and firewalls.

8.6.7 *Network Integration – Consolidation and Virtualization*

The consolidation and virtualization component helps clients understand, plan for, and meet the new demands of virtualized servers and storage while also addressing consolidation and virtualization of the network itself to further reduce infrastructure costs.

- Designed to address the new demands that virtualized servers and storage devices place on the network *and* the benefits of consolidating and virtualizing the network itself.
- Deliver cost-effective, optimized networking infrastructures that support and fully contribute to a responsive, consolidated, and virtualized IT environment.

Cloud vendors can help businesses migrate from a traditional, isolated, static network design for the datacenter to one that is integrated with other IT resources to provide dynamic, scalable network resources. Regardless of the brands of technology in their environment, network consolidation and virtualization services are designed to offer guidance throughout the process – from developing the strategy and assessing the current infrastructure to designing and implementing a networking infrastructure that comprehensively supports a dynamic infrastructure.

In the big picture, IT infrastructure is moving left to right . . . servers, storage, and networking are becoming more and more interdependent, leading ultimately to a set of resources that are provisioned together to deliver services:

- **Legacy Environment:** Static, endpoint agnostic, strict, limited change windows, proliferation of special-purpose devices (firewalls, load balancers, IPS).
- **Device Virtualization:** Physical consolidation and optimization, basic virtualization of servers, storage, and network, simply network management.

- **System Virtualization:** Connecting virtualized servers and storage, support platform-specific network requirements, multiple layers of network virtualization.
- **Cloud Computing:** Architect responsive, secure network, support automated provisioning of servers, storage, and network; increase operational savings.

8.6.8 Datacenter Network Thinking Has to Change

Static, secure datacenter networks that meet their non-functional requirements through limited, controlled changes are no longer adequate. Datacenter networks must become significantly more flexible and responsive, capable of dynamic change. The datacenter network is a critical success factor for storage and server virtualization initiatives – ignore it at your peril. Consolidation and virtualization in the datacenter is increasing demands on the network in terms of throughput and traffic patterns, upending traditional performance and capacity 'rules of thumb'. Virtual Machine (VM) mobility and mixed platform environments will require the network to be much more dynamic in terms of scaling and services provided, to transfer workloads without disruption to end-users and business processes. The network must be integrated into the overall IT systems management environment to provide dynamic services in response to automated provisioning.

8.7 SUMMARY

This chapter provides an overview of the architectures, features, and benefits of cloud infrastructure solutions which provide a cost-effective solution for any datacenter cloud implementation.

CHAPTER 9

CLOUD AND SOA

Introduction

SOA Journey to Infrastructure

SOA and Cloud

SOA Defined

SOA and IAAS

SOA-Based Cloud Infrastructure Steps

SOA Business and IT Services

Summary

9.1 INTRODUCTION

Any enterprise-wide transformation has significant challenges for people, processes, and technology. Therefore, identifying the challenges ahead of time and defining a mitigated approach can help such a transformation succeed. Some of the challenges include resistance to change by people and organizations. Factors include roles and responsibilities, management, skills development and discipline, and cultural shifts. It also includes creating awareness in the organization for the need to drive such a transformation in the best interests of business. It is very difficult to deal with infrastructure complexity, including hardware, software, and applications across disparate environments ('line of business' stakeholders, partners, and customers). Well-planned assessments are needed to understand where to start and how to progress in a staged way.

Service management is one of the similarities between cloud infrastructure and SOA approaches. Developing an integrated service management approach for both the application services and infrastructure services together will drive efficiency in IT operations by improving resource utilization and improving service levels. Such an integrated service management can move IT towards an end-to-end service-oriented environment. Such an environment will enable business agility by better aligning IT with the business.

9.1.1 Enterprise Infrastructure and SOA

Design and provisioning of enterprise infrastructure must be focussed on the needs of enterprise organizations. Through comprehensive capabilities of products, services, and integrated solutions, IT organizations are delivering increasingly complex solutions, driven by Service-Oriented Architecture (SOA) and related methodologies.

Achieving the ambitious growth goals that characterize leading IT organizations requires continuing investments in information technology, taking advantage of emerging capabilities. A future technology platform will need to support agile business organizations through the simplification of information systems and reduce the complexity of the IT ecosystem through consolidation and rationalization.

We need a governance model for a heterogeneous environment owned by many parties and providing end-to-end IT infrastructure. This governance model will define the IT infrastructure service requirements in support of an integrated service offering for business systems.

As SOA projects are deployed, effective design of supporting infrastructure becomes critical. SOA introduces requirements for availability, service continuity, monitoring, scalability, and geographic dispersion that are different than those of past architectures.

SOA makes IT applications into composite applications. Instead of traditional monolithic applications, composite applications are created, composed of many services often developed and deployed independently by separate development teams on different schedules. By adhering to common standards and interfaces, development of new composite applications and extension of existing applications are made easier through the reuse of existing services and the rapid integration of new services.

Similar concepts to SOA drive cloud infrastructure, an approach that makes IT infrastructure a collection of service components with common standards and interfaces. Cloud infrastructure makes the deployment of new infrastructure and the extension of existing infrastructure easier through the reuse of existing services and the rapid integration of new services.

Cloud infrastructure service components include physical infrastructure (such as processors, memory, storage, and I/O networks), system software (firmware, operating systems), and management software (monitoring, provisioning, workload management).

While cloud infrastructure is particularly suited to support SOA applications, Service Oriented Infrastructure (SOI) is also well-suited to legacy application support. The service components of cloud infrastructure are independent of application architecture and are capable to providing flexible support to any application.

Cloud infrastructure strongly leverages virtualization technologies, which enables rapid deployment and redeployment of service components.

9.2 SOA JOURNEY TO INFRASTRUCTURE

The path to transformation consists of a long journey with a staged approach, leading to the ultimate goal of a service-oriented enterprise. Multiple islands of disparate infrastructures in today's environment need to be consolidated to gain control, reduce costs, and become operationally efficient. The next step is to introduce virtualized infrastructure to improve utilization levels and allow dynamic flexibility to move resources and capacity to meet fluctuating workload demands. It is important to note how the service orientation can be achieved by building capabilities on top of virtualized and automated infrastructure. Service orientation is a state where infrastructure is provided and utilized as a service, rather than in piecemeal. Latest innovations such as cloud computing will help to further expand the service-oriented paradigm, to meet the scaling demands of future state of businesses.

9.3 SOA AND CLOUD

SOA binds how you will both deliver and leverage cloud-based services. Cloud computing relies on service-orientation (virtualization at the application layer) to loosely couple applications to the underlying infrastructure model for using Web services – service requestors, service registry, service providers. It uses Web services to compose complex, customizable, distributed applications and encapsulate legacy applications. It helps organize stove-piped applications into collective integrated services for interoperability and extensibility. SOA serves as the foundation for the move into cloud computing and it owns the characteristics of a cloud including a shared infrastructure, self-service capabilities, and the fact that it will be virtualized (Figure 9.1).

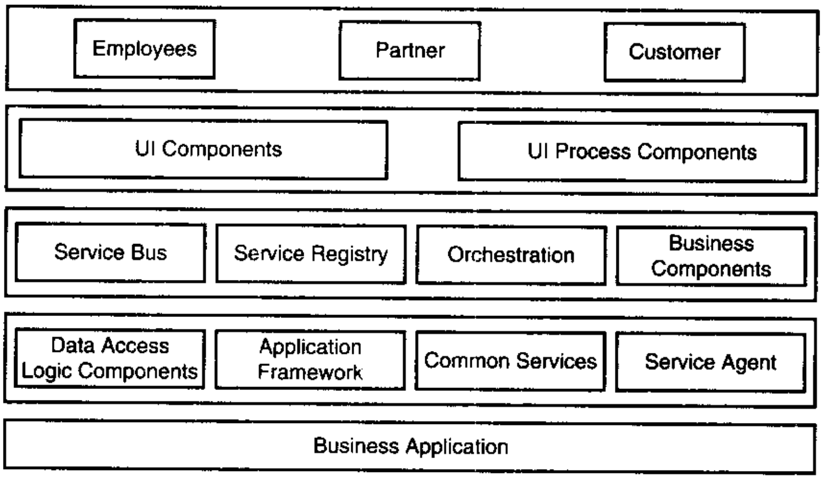

FIGURE 9.1 SOA Model.

Cloud computing is an infrastructure management and services deployment method with virtualized resources and it is managed as a single large resource. Clouds share and leverage characteristics of SOA with flexibility and agility. Applications and services reused in new and dynamic ways and rapid deployment happens in SOA-based cloud implementation.

SOA infrastructure is required in order to effectively apply service orientation to an enterprise or large-scale software component development. Basically, SOA infrastructure consists of middleware, physical infrastructure, and management – all together covering the non-functional requirements.

Service-oriented architecture is an application framework for building better applications. Web services and SOA do not replace applications as they are today; they complement their functionality and allow for better reuse and business flexibility. SOA helps break an application framework into discrete service components (that is, mini-applications) so they can be reused as common services between different applications.

IT infrastructure must continue to evolve and mature to support the new demands on a distributed and virtualized application framework. SOA applications require the same end-to-end performance, security, and management.

The key to SOA infrastructure understands what will change in the environment and what tools are available for infrastructure architecture and design. The best SOA infrastructures have been designed with both application and infrastructure perspective in mind.

A well-designed SOA infrastructure is a mix of current and SOA infrastructure technologies. SOA and traditional applications don't exist outside of one another. Applications are all a part of a shared services environment and use common infrastructure components. Traditional system designs need to be updated to support the new application requirements.

Clouds enable deployment of cloud services, and SOA is the most sophisticated architectural approach for the building and delivery of services. SOA is a design pattern that is composed of loosely coupled, discoverable, reusable, interoperable platform-neutral services. Each of these services follows a well-defined standard, and can be bound or unbound at any time and as needed. The value of SOA comes from having an architecture that readily accommodates change.

Clouds are about infrastructure and deployment technology. The concept of service delivery is independent of deployment scale, and may operate via a connection between only two or three computers, whereas cloud computing represents a much larger-scale implementation. SOA is about proper system architecture, and even an excellent infrastructure can't save a bad architecture. SOA is the way an enterprise builds, maintains, governs, and orchestrates the services you deliver; cloud computing is an instance of SOA. SaaS is a term used by a group of companies or individuals to mean that they are hosting a set of software services over the Web. SaaS focuses on software hosted as a service, and may be considered as a consumption model in which a user is involved. SOA focuses on software designed as a service, and is a design model in which there is no restriction on the consumer.

Private clouds can be seen as simply SOAs with the addition of virtualization and self-provisioning. Those who use private clouds, or virtualization, typically break down new and old applications as services, processes, and data and address each as an architectural component that may be freely distributed in the private clouds.

SOA is important to cloud computing, and the use of SOA will promote the adoption of cloud computing. Enabling SOA is important for enabling an infrastructure for service delivery – a cloud. Most experts agree that without SOA, a move to clouds will be tough to justify financially, because it will cost too much to re-engineer legacy systems that are not built to be exposed outside of the usual user community. Enterprise cloud initiatives require decoupled data systems working together – without the need for personnel and other resources to set up and maintain them – integration is key. The loosely coupled aspect of SOA is very important.

The best scenario for moving services to the cloud is when applications, processes, and data are more loosely coupled and less dependent on each other. Companies who aren't practicing SOA will be tightly coupled to their databases and to their infrastructure, making it very hard for them to move, or shift, or change things around.

When a cloud user initiates a service, it calls a mechanism to expose legacy functionality as service on the cloud. This can require integration across firewalls and across technology boundaries. An enterprise service bus by definition is equipped to provide this capability and this becomes a vital component of the IT infrastructure that leverages cloud computing. With SOA, an enterprise can look at their entire offering and decide to move certain pieces into cloud, and not other pieces. Without SOA, it almost becomes an all or nothing proposition – not a recipe for success.

A 'service communications backbone' is needed to run between the different clouds being used, which will allow users to utilize remote services from any cloud without having to deal with connectivity and interoperability issues. It is a simple concept, but without it, cloud-to-cloud interoperability issues may limit the growth of cloud computing. This is really going to require state-of-the-science SOA, with the ability to access thousands of services that could be hosted anywhere and to abstract from the interoperability issues. As is always the case with such industry efforts, the standards process takes time. The trick is to develop service architectures that won't require an overhaul in the future based on specs yet to be defined.

9.3.1 *Infrastructure Technologies*

Cloud infrastructure is based on virtualization – dynamic systems that enable the definition and delivery of resources on demand. Current server technology can deliver hundreds of virtual servers on small cluster of physical servers, enabling flexibility and high availability.

In a virtual environment, workloads can be moved dynamically between components, allowing minimal unplanned downtime and no planned downtime. Each server contains a pool of processor, memory, and I/O resources that can be dynamically assigned and reassigned to meet needs. Surplus capacity can be pre-provisioned, at no cost until activated.

9.4 SOA DEFINED

SOA is an approach to architecture that is intended to promote flexibility through encapsulation and loose coupling. SOA functions are defined and exposed as 'services' and there is only one instance of each service implementation, either at each service, for example exchange rate calculation, is deployed in one place and one place only, and is remotely invoked by anything that needs to use it. Deployment time is less as each service is built once, but re-deployed to be invoked semi-locally wherever it is needed.

SOA is about an evolving living organism and not about building a house. This is an ongoing journey and not a project that finishes with a concrete result. Agility for the business is an important factor for business continuity as it helps faster solutions to changing business priorities and leverages the competitive effectiveness of business change requirements.

SOA is defined by what a service is. Services are defined by the following characteristics:

- Explicit, implementation-independent interfaces.
- Loosely bound.
- Invoked through communication protocols.
- Stress location transparency and interoperability.
- Encapsulate reusable business function.

Conceptually, SOA can be visualized by the roles of the individuals in any organization. The architect sees SOA from the perspective of the entire business and uses SOA implementation to bridge the gaps of the business.

SOA is very flexible; therefore, it facilitates the different elements of business. The most important characteristic of SOA is the flexibility to treat elements of business like:

- Business processes.
- Underlying IT infrastructure.
- Secure standardized components (services).
- Changing business priorities.

So when we look at the SOA vision we need to look at three aspects:

- The business view of a service – what is needed to support the business process.
- The architecture view of a service – how do we define and design these services.
- The implementation view of a service – how do we implement the service through component deployed on the technical infrastructure?

In order to run the business in a smoother way, we will have to bundle the business requirements in a simplistic way and it should be standardized. This creates the service offerings and helps to get the right information from the right source particularly information about when it is needed. This enables us to reuse and combine different other service offerings to answer the requirements of winning against competition.

9.4.1 SOA Lifecycle

The SOA lifecycle not only resembles 'traditional' application lifecycles, but also introduces new terminology. SOA in terms of a lifecycle starts in the SOA Model phase where organizations gather business requirements and information about designing their business processes. Once they have optimized the business processes, they implement it by combining new and existing services. The assets are then deployed into a secure and integrated environment for integrating people, processes, and information. Once deployed, customers manage and monitor from both an IT and a business perspective. Information gathered during the Manage phase is fed back into the lifecycle for continuous process improvement. Underpinning all of these lifecycle stages is governance, which provides guidance and oversight for the SOA project.

9.4.2 Service-Oriented Computing

Service-orientation is a design paradigm comprising a specific set of design principles. Its most important feature is its reliance of the 'separation of concerns' design philosophy. Separation of concern is based on the simple fact that a problem becomes easier to approach if it is divided up and handled separately.

The first question that should come into the mind is what is a service. Service is not only limited to the software or Information technology, actually it is culture of the organization and how it performs its entire operations on a day-to-day basis. We can divide all these tasks into small processes and investigate the processes that are repeatable and can be used as business continuity process. This also implements the agility factor for the business. Now, if we talk about service orientation, it is based on the integration of all the business processes as related processes to get the achievable outcomes intended from the business.

Next comes the technology associated with SOA. This visualizes the architectural aspects of the service orientation to make the process simple and gives the option of composite application. The composite application ties the running process and business requirements in such a way that it helps to achieve the business goals.

9.5 SOA AND IAAS

Major industry analysts view cloud infrastructure as a key IT ingredient for business agility. With a predicted 60 percent of IT spending being applied to infrastructure, analysts recommend an IT Infrastructure that is:

- Shared across customers, business units, and applications.
- Dynamically driven by business policies and service level requirements.

The analysts view IT Virtualization and IT Automation as two major elements in realizing infrastructure as service.

IT Virtualization is viewed as a technological aspect of cloud infrastructure in order to create a pool of infrastructure resources, such as computing power and data storage, in order to mask the physical nature of the boundaries from the users. In other words, virtualized resources are viewed as fluid utility services for the consumers to consume as needed (Figure 9.2).

IT Automation, on the other hand, is viewed as a way to better govern the utility model infrastructure services, enabling policy-based, service-oriented, dynamic management of underlying virtualized resources. The recommended implementation approach towards SOA looks towards a strategic return on investment, rather than a quick fix, tactical return.

9.5.1 Architecture

Cloud infrastructure has many service components. However, they need not all be implemented concurrently. Services can be divided into four domains: Application Services, Information Services, Common IT Services, and Infrastructure Services. Within each domain, SOA can be measured and charted across a continuum of increasing dynamism and partner involvement.

Application Services provide the application frameworks to enhance the execution of business services through software engineering. Adopting new technologies and techniques can accelerate the delivery of new services through the use of consistent, repeatable service-oriented architectures.

Information Services provide a common, repeatable method for cataloguing, accessing, and managing information. Innovative technologies can streamline information access and data management, making it easier to integrate packages and new acquisitions. Common IT Services create enterprise pools of commonly used IT services. Simplifying the environment can enhance management and cost and increase responsiveness.

Infrastructure Services provide pools of processing and networking resources for applications and business functions. Today, these resources may be isolated into business silos, but with virtualization, they can evolve into virtual pools that are dynamically allocated based on business need.

Continuous improvement is mapped within the maturity levels of the company itself and can measured in each domain of the service. SOA plays a fruitful exercise to decide on how to implement the design the infrastructure based on SOA principles to attain the targeted goals. It offers a number of business values.

Business Agility

- This helps in defining the right time to launch or rapidly scale the deployment efforts needed to implement the new solutions.

Lower Cost of Operations

- This helps in utilizing the virtual pools efficiently which decreases the chance of procuring the new systems.
- It helps in increasing the overall effectiveness when we work in an automated environment.

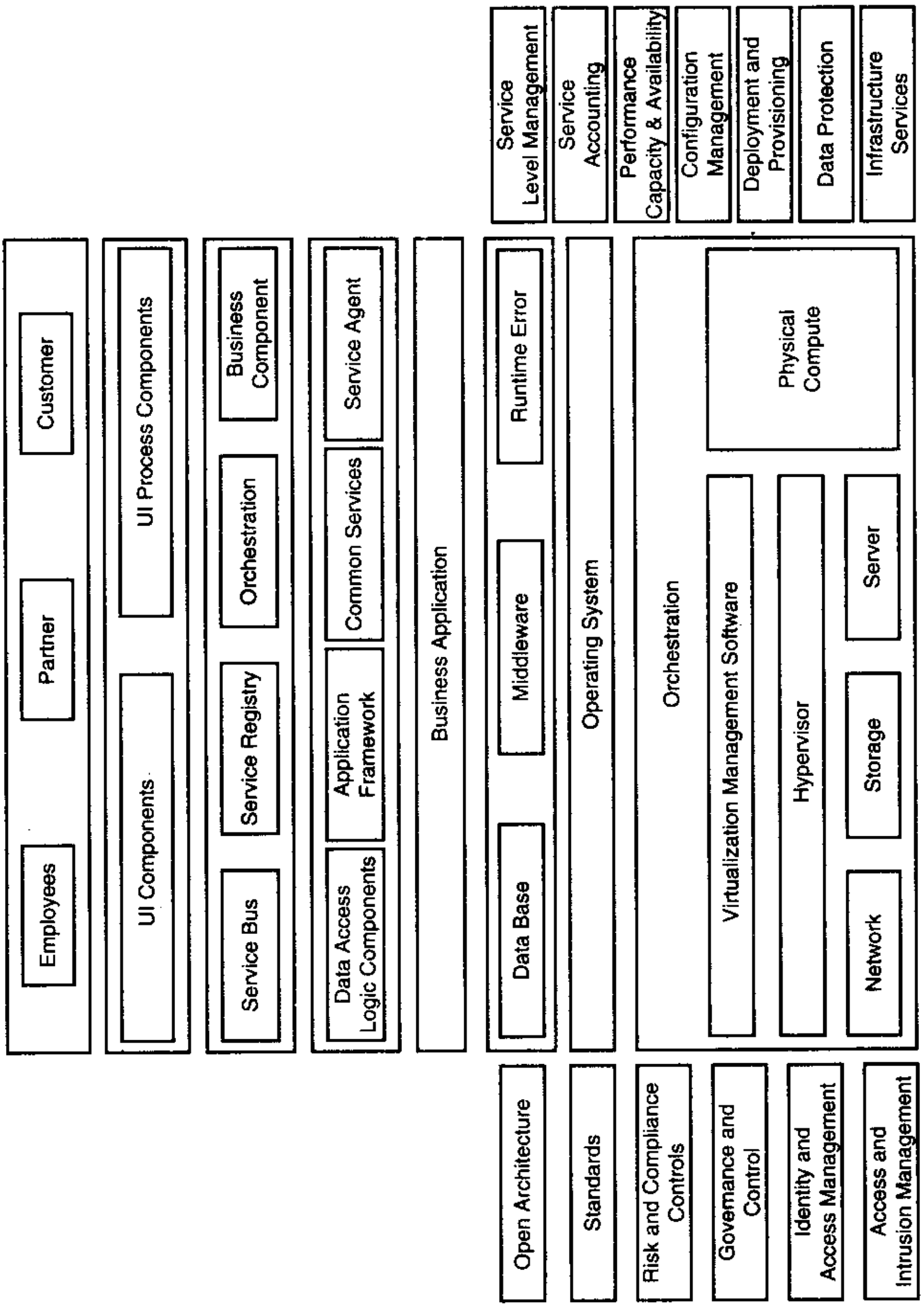

FIGURE 9.2 Cloud IT Service Management.

Improved Service Levels

- SOA-based infrastructure helps to adhere to the SLAs efficiently and helps in orchestrating the resource as per the rules and policies.
- Business analytics helps to decide different predictive, proactive, alterative approaches when we follow SOA.

Efficient Information Management

Efficient centralized virtual environment enables:

- Information dissemination.
- Data replication.
- Business continuity protection.
- Regulatory compliance.
- Maximizing resource utilization.
- Rapid deployment of new applications.
- More timely response to changing business conditions.

Regulatory Compliance

- We need monitoring to track the performance of the services to confirm that they comply with the regulatory compliances. Workflow automation helps for the same.
- It is important to adhere with energy emission regulations and efficiencies to adopt Greener solutions.
- If we have centralized data storage it will help the audit process. Centralized data storage facilitates and accelerates audits.

Energy Efficiency

Energy requirements for the datacenter are rising day by day. To match the requirements we have the answer: SOA-based cloud infrastructure. SOA-based cloud infrastructure substantially improves:

- Computing.
- Storage.
- Network utilization.
- Datacenter energy efficiency.

9.6 SOA-BASED CLOUD INFRASTRUCTURE STEPS

Organizations intent upon leveraging cloud infrastructure should consider the following steps:

- **Analysis and Strategy:** It is recommended to have an incremental, phased approach for adopting SOA and cloud infrastructure. A good starting point is to conduct a Business Innovation Assessment to identify business needs and key areas for impacts, and use them to develop business value cases for SOA adoption. It is also recommended to conduct an Enterprise IT Architecture Assessment to determine the IT readiness, including the applications and integration capabilities to support the business needs and the current gaps. Furthermore, an IT Infrastructure Assessment should be conducted to determine capabilities and entry points for Service-Oriented Infrastructure.

- **Planning:** Once the business needs and IT gaps are identified, a strategic and tactical planning effort can be launched to develop an IT Value Case and Roadmap for incremental IT transformation to enable business innovation/agility leveraging SOA. Through a careful portfolio analysis and change management process, the current IT efforts can be aligned and adjusted and tactical/focused projects can be selected based on the strategy and roadmap. At the same time, organizational impacts should be analyzed and a SOA governance model, standards, and guiding principles should be developed to accelerate and manage the pace of the transformation.
- **Implementation:** It is an excellent idea to establish an enterprise-level SOA as well as the development and run-time environment standards before the first set of SOA projects are launched. It is best to couple the virtualization projects with SOA so that the benefits of the virtualization can be realized at the lower level (service-level versus server-level). SOA and cloud infrastructure governance should be incorporated into existing IT governance bodies and performance and service level agreements impacts should be monitored and managed under the same set of business rules.
- **Value-Driven:** It is important to note that the purposes of SOA and cloud infrastructure are to improve the business performance, flexibility, and agility so that IT can support business at the business speed. All SOA projects should be driven based on business value rather than technical merits alone. Technical merits can only be realized when they match the business needs.

9.6.1 SOA and Cloud Infrastructure

SOA is an approach to decompose business processes and applications into loosely coupled components of service providers and consumers and then connects them through enterprise service bus. It enables enterprises to reuse existing business and IT components quickly to develop new capabilities and software solutions. SOA provides an enabling foundation for enhancing business and IT flexibility and agility. The approach is gaining significant momentum, not only in designing new solutions, but also transforming monolithically defined legacy applications.

We can bank on the traditional IT model that was proved successful for the deployments. These models can be based on historical data processing of data and transactions. As these are established models, they gel well with the high structured environment. But they breaks down when you try to extend it into applications or processes that aren't so highly structured. They either feel too complex and static (long-term ERP projects, for example) or are plain impossible.

With the advent of Web and Internet, we got the new wave itself for new paradigm of models. It is based on open standards and linked easily for different requirements and components, which you can then use for relatively simple activities like communications, browsing, searching, and sending e-mail. It works incredibly well. But it soon became clear that the Internet standards and mechanisms were needed to be extended to handle more sophisticated applications.

The SOA-based cloud computing model builds on the IT and Internet models. It is based on what we call a service-oriented architecture, which essentially provides us with a set of modular components to be defined and manipulated (Web services), and a set of XML-based standards for doing so. Since the characteristics of the components can now be expressed in XML, we can define applications that work and manipulate these modular components. It enables a much more flexible and real-time way of implementing business policies than was possible with more structured computing models.

SOA-based cloud computing is not about technology for the sake of technology – it is about enabling new ways of doing business. It is about helping a company to reach new levels of maturity while continuing to deliver the best in class services with productivity; these are necessary to improve the bottom line.

As the processes are integrated in SOA environment it gives an option to the enterprise to deal with any type of situation and answer any type of customer demand with the help of partners, suppliers, and customers.

One should follow the most important approach for SOA to consider the business' underlying principle and not only the technical foundation of the business as it will help to determine the cost of investment. Now we are in the era where models are evolving and changes are very dynamic; therefore, all the technological steps should have business backup to support it. We can consider the different types of capabilities related to organization, strategic values, and market factors that are driving the business.

SOA means a company finds a pragmatic balance between technical rigor and time-to-market. IT organizations realize that perfect technology stacks cost far more than they deliver, but they also recognize that delivering key functions and reliability of service make it easy for employees and customers to use the software – and drive the desired business results.

When we adopt SOA, ROI is required to gauge the return on investment to understand the value of the investment model. Simplistic approach will be to calculate the cost of change and see that it should not be more than the actual cost of implementation and highlights to adopt the changed approach.

Diversity in the portfolio gives the alternative to face the risks, it is better to work for a different set of application and not on the single application. Therefore, it is recommended to give space to larger range of future prospects and create value to improve the business.

It is important to integrate and control the business functions across all the units of business with the help of partners and customers. So automation in the process delivery should be valued as it increases the transparency in the system and reduces the manual intervention also. Sometimes it alerts about opportunities to grab and get unexpected results out of them.

Competitive advantage and system-based performance are the levers to progress for the business. But the balance between the two is very important. SOA helps to balance both. It helps to consider the risks and overcome with spikes in standards and product deployments. This ensures security and companies can leverage the power of software application in a more efficient way.

9.7 SOA BUSINESS AND IT SERVICES

We need different management tools for SOA for comprehensive integration of the SOA. These tools help to leverage the benefits of infrastructure services. They also help to measure the performance of business functions and manage the run-time application and system across the portfolio of business functions. These development tools aid to get the specific outcome based on the skills and roles possessed by the people in any organization.

Business analysts who analyze business process requirements need modelling tools to chart and simulate business processes. Software architects need tool perspectives to model data, functional flows, system interactions, etc. Integration specialists require capabilities to configure specific interconnections in the integration solution. Programmers need tools to develop new business logic with little concern for the underlying platform. When we follow the SOA implementation, it is evident that the persons in the organization will use the systems based on their roles in the organization. The tool environment and deployment framework allows working in an integrated way and using the development tools in a joint manner that ensures collaboration and asset management. So the activities, like using the tools such as version control and project management tools, are provided in the SOA framework under the banner of unified development platform. Generating metrics to measure the performance is pivotal in the progress of the business. SOA services incorporate functions to generate the metrics to control the system and processes. This requires the special competencies to deal with the IT operations with the experience skill sets of business analysts and professional of an enterprise These capabilities are delivered through a set of comprehensive services that collect and display IT and process-level data, so that business dashboards, administrative dashboards, and other IT level displays can manage system resources and business processes. These tools make it possible for LOB and IT personnel to determine business process paths that may be inefficient problems in specific processes or the relationship of system performance to business process performance so that IT personnel and assets are tied more directly to the business success of the enterprise.

9.8 SUMMARY

This chapter shows the integration of SOA with cloud technology. SOA is essentially the idea that companies can treat their applications and processes as defined components that can be mixed and matched at will. SOA is much more the architecture flavour of the day – in fact, it's a new way of thinking about enabling cloud technology.

CHAPTER 10

CLOUD MOBILITY

10.1 INTRODUCTION

The rising utilization of consumer applications, technologies, and utilization paradigm is continuing to shape user expectations about the workstyle milieu. Many workers are interested in carrying their own device to their offices. Workers want to be able to perform their jobs using the platforms, applications, online tools, and services they choose. 'Bring your own device' (BYOD) enables workers to choose the platforms and devices that best fit their needs, providing them with greater flexibility and ultimately making them more productive. IT consumerization, especially the desire for BYOD, is a significant trend that both offers benefits and incurs expenses. The benefits include the following:

- Enhanced employee productivity and job satisfaction.
- Reduced cost to the company compared to the cost of providing the devices.
- Greater business agility gained by the use of a wider array of usage models, many of which offer a greater degree of mobility. The expenses associated with BYODs include those related to developing new infrastructure and implementing the controls required to bolster back-end device support because of potential information security risks. However, the benefits far outweigh the costs.

Companies are using mobile applications in greater-than-ever numbers to get better business dexterity and deliver greater customer values. Information seized and distributed at the boundary of the companies broadens the accomplishment of enterprise applications to major decision support systems. Previously, mobile communications for an enterprise tended to be isolated and all over the place. To productively distribute on the idea of mobility and expand the reach of mobility, enterprises require a structural design and approach that allow a comprehensive view and integrate mobility as an fundamental fraction of the enterprise architecture.

With the amplified espousal of mobility as a premeditated plan, these next generation strategic architectural advancements for mobility will enable companies to solve their mobility requirements for diverse business problems across multi-tenant environment in an enterprise.

10.2 THE BUSINESS PROBLEM

Enterprises have spent the last many years selecting best of a variety of software and applications to rationalize their business processes and their support systems. These have ranged from Enterprise Resource Planning (ERP) systems to tradition applications and business intelligence platforms. A lot of attempt has also been made to put together various applications that often have diverse data formats and semantic dissimilarities.

Enterprises have come to comprehend that the next step in the fruition of their business progression is to mobilize key business elements that will provide both domestic and peripheral interactions via mobile instruments. These devices are not just mobiles, and include various devices like tablet phones, iPads, ipods, notes, etc. and any prospective devices that can communicate with the existing applications.

In order to move towards a mobilized world, enterprises will have to struggle with several issues which we discuss in the following sections:

10.2.1 Segregate Systems/Data and Intangible Business Processes

Every enterprise has multi-tenancy, that have their own applications, processes, and data. It is essential to be able to segregate the processes from each other, so that transformations in the department-specific processes do not affect the whole enterprise. This can be enabled only if a general service level is defined for all the business processes that offer a basic edge and a conceptual layer to incorporate with every persona-based process. Any mobility response needs to present a methodology, where any business process that one can wish to mobilize is understood first and the interfaces are definite with an understand able generalization. The service levels thus developed will facilitate external devices and applications to incorporate safely, and any internal transform will not influence or split the integration. The other key benefit of this methodology is that only the smallest amount of requisite functionality is provided to outside applications to be able to incorporate into a precise business process. Conception and segregation are two foundation stones of good design and allow wider usability of any organization applications in the centre of varying business requirements which austerely affect the general amalgamation.

10.2.2 Security and Access Controls

Data security is of dominant significance in the overall system of any mobile solution — securities like authorization, authentication, and data encryption are obligatory. It is imperative for the solution to clearly endow with for authorized access not only to dissimilar applications but also at a more coarse level, that is data within an application.

10.2.3 Amalgamation

No application can exist in segregation — enterprise-level systems are not only inter-connected but also amalgamate through data. As data flow from one system to a different system to help complete business processes, data reliability and validation becomes significant to business compliancy. As a consequence, the solution should be able to represent methodology that allows standards-based amalgamation and integrations to enterprise resource systems and even custom applications through standard and open amalgamation methodologies.

10.2.4 Elasticity

The number of mobile devices is increasing more and more and enterprises have come to recognize that with time, an ever-increasing number of their accessible processes will have to be available on mobile devices. The mobility solution should be elastic to cater to increasing devices; this will need to be made accessible over time.

10.2.5 Support

Mobile devices come with different functionalities, form factors, abilities, and in different types. And it is not just about a mobile phone anymore — there are mobile phones, tablets, and various other inter-connected devices that will need to put together into the solution. All well-liked mobile device OS should be supported. The mobility solution be able to allow right of entry or amalgamation by various mobile devices via an Open Standards Service Level. This Service Level should be idyllically written once to accommodate manifold applications from numerous devices.

10.2.6 *Infrastructure*

Enterprises need spotlight on their business functionalities relatively more than on their infrastructure services, as it is the second step. The mobility solution should be able to make available various infrastructure services like Service management, Monitoring, Scheduling Engine, Allocation Engine, Notification Engine, Resource Integration Engines, Integration Adapters, Identity and Access Management, etc.

10.3 MOBILE ENTERPRISE APPLICATION PLATFORMS

Mobile Enterprise Application Platforms speed up and make simpler the development, use, and administration of smart mobile devices. They endow with a set of toolkits and a coupled runtime infrastructure to connect mobile task workers to a variety of data foundries in a device and network without any problem.

Mobile Enterprise Application Platforms promote 'any mobile application to any device' model strategy that brings together five key elements - mobile devices, middleware, management toolkits, application development environment, and resource integration framework. Together, they work with each other and the on-premise infrastructure for uninterrupted mobile connections and networks. By bringing these advantages to mobile application development and implementation, Mobile Enterprise Application Platforms allow companies to fruitfully address many of the challenges faced today by mobile developers.

10.3.1 *Freedom of Choice*

With a distributed development methodology that supports diverse device types and platforms, Mobile Enterprise Application Platforms can get rid of the recurring, resource-intensive activities involved in developing and implementing mobile applications. This enables enterprises to lucratively flick the application development proportion, so they can spend less time on familiarizing applications for devices. Instead, they can use the efforts on developing mobile applications that bring value to the customer business and a pertinent, appealing experience to end-users.

Developers can also contemplate on building powerful business layers and content-rich interfaces that acclimatize to different user needs. That's where the added value to users comes in: With a distributed development methodology that supports diverse device types and platforms, Mobile Enterprise Application Platforms can get rid of the recurring, resource-intensive activities involved in developing and implementing mobile applications. This gives developers the litheness to support different workflows within the application lifecycle.

10.3.2 Agility

Mobile Enterprise Application Platforms solutions provide an end-to-end development platform for developing, designing, testing, and building mobile applications across heterogeneous devices and OS platforms. Developers can also contemplate on building powerful business layers and content-rich interfaces that acclimatize to different user needs. That's where the added value to users comes in: With a distributed development methodology that supports diverse device types and platforms, Mobile Enterprise Application Platforms can get rid of the recurring, resource-intensive activities involved in developing and implementing mobile applications. This give developers the litheness to support different workflows within the application lifecycle environment uses popular open source IDEs.

This elastic platform is vital for enterprises that want to accelerate development while lowering costs, because it permits developers to utilize existing expertise and experience. To make development simpler, Mobile Enterprise Application Platforms also comprise templates and baseline application stacks; tools to develop, test, and debug, simulate an application on a device emulator, which considerably speeds up development and testing of devices.

10.3.3 Feature Rich

Applications developed in a Mobile Enterprise Application Platforms permits users to take advantage of the unique capabilities of their selected mobile devices, as well as consumables, like barcode scanners and printers.

10.3.4 Robust Connectivity

Mobile middleware is at the focal point of the Mobile Enterprise Application Platforms architecture. Acting as a controller for bi-directional communication between backend infra and development systems and mobile devices, this is where the nucleus wireless message change takes place, as also transaction routing functions.

Mobile Enterprise Application Platforms connects with data sources to mine, change, and assimilate data. It then encrypts the data and propels back in real-time to the mobile middleware.

A Mobile Enterprise Application Platforms integration methodology can comprise a range of pre-built application integrators for adaptors with bundled and homegrown applications running on enterprise management systems. This decreases the need to build up integrators from the beginning or to modify existing integrators.

10.3.5 Off-line On-premise Integration to Business Processes with the Clients

In a Mobile Enterprise Application Platforms, applications can work separately of a master server connection, so users can keep on working offline. Applications run in the vicinity on devices for better response time, and updates are pushed without human intervention when the mobile devices reconnect to the network.

10.4 MOBILE APPLICATION ARCHITECTURE OVERVIEW

Most mobility systems widen an existing business system or edge with an existing system. There are typically three major parts to a mobile architecture: An enterprise system, a middleware, a handheld App (Figure 10.1).

The rationale for middleware that is generally desirable, is to grant data changes, apply business intelligence, and be a focal point of messaging for the devices. If a new business system is being architected or customized then no middleware may be essential; the appropriate flow can be built into the system to converse with the devices from the beginning. Most business systems are not written from the scratch very often and it is expensively impractical to rewrite them just to provide mobility to the application. It also works like a management server. Mobile system developments frequently engage various technologies due to environment limitations.

10.4.1 Device Application Installations

If the application is being installed on new mobile devices, there is a good probability that some arrangement will be required. After a device is started, the installed application must be loaded. Various companies use IP-based mechanism for installing applications and configurations.

10.4.2 Upgrades

There are software updates that administer device configurations. This management pack works on a client that sits on the device. Manufacturers' IP-based packages must classically be written for the administration that identify the software and upgrade files to be downloaded and installed.

10.4.3 User Interface

It is extremely difficult to design the graphical user interface on mobility-based devices as there is usually very little screen space and very small keypad for data entry. If the application or data is multi-faceted, the user will be required to act together, that is application and data, with many screen items. Intricate screens will need to be separated and divided into subscreens or tabbed icons.

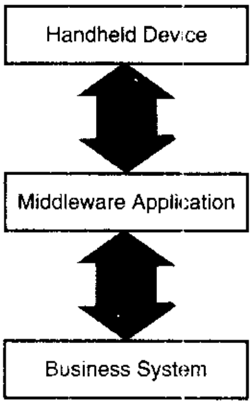

FIGURE 10.1 Mobility application model.

10.4.4 *Performance*

Servers and desktop have improved considerably and now their performance is characteristically not a matter of discussion. Mobile-based computing devices, however, require different treatment as these devices are very slow. Intricate user interface, CPU exhaustive algorithms, and data processing can straight forwardly make an application user aggressive. It is important to take corrective measures to evade performance drawbacks.

10.4.5 *Memory Management*

Most mobile computing devices come with a limited memory even though memory is becoming cheaper with each passing day. It is really difficult to push configuration changes and upgrade the patches. Data storage is also a critical factor as it is very limited in handheld devices.

10.4.6 *Security*

As mobile devices are becoming computing devices, security is of prime concern, for example, security of desktops and servers. If the device is lost there should be provision to encrypt the data so that it cannot be used by those who have stolen it. Some of the devices come with various methods of authentication like biometrics, voice recognition, login credentials, etc. (Figure 10.2).

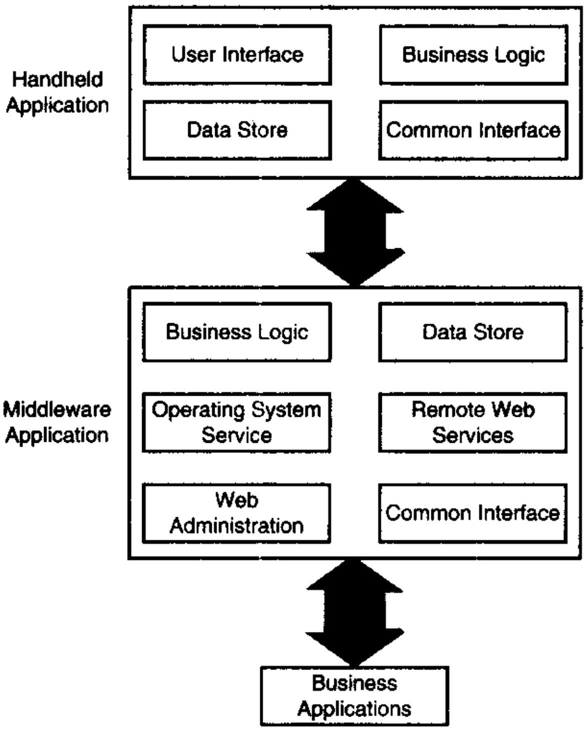

FIGURE 10.2 Mobility application architecture.

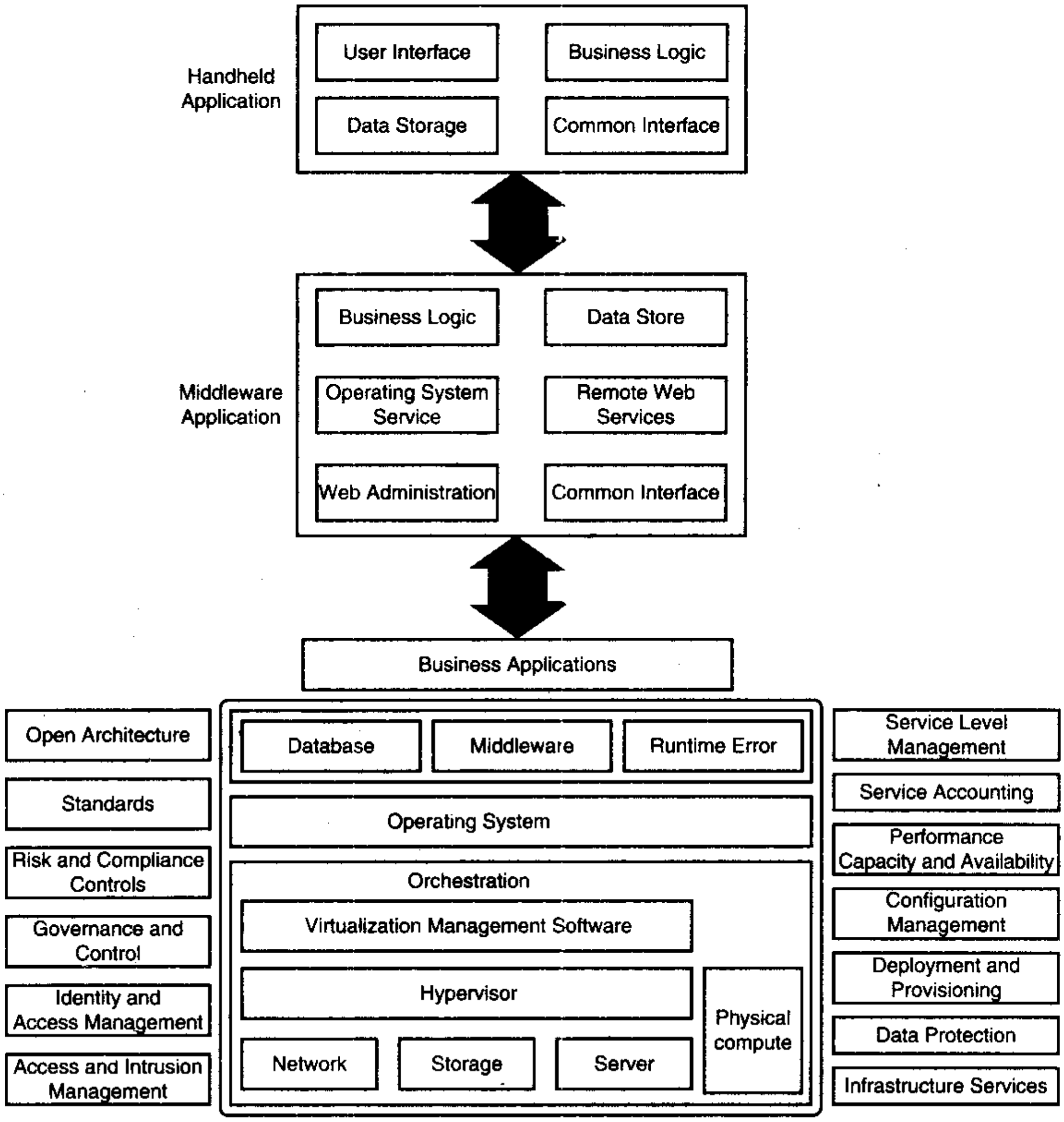

FIGURE 10.3 Mobility cloud model.

10.4.7 Business System

The business system can be platform running an enterprise management system. The only crossing point is the import/export interface already available in enterprise server. Data from the handheld mobility devices must be imported by hand manually or automatically in the business system.

10.4.8 Middleware Application

The middleware application can be included in OS service, a web service, and a multi-threaded socket server. It is accountable for downloading business data from the server and changing this data into smaller files that are better organized for the handheld mobile computing devices. It also provides the data files to mobile-based applications of handheld devices through server. Also it retrieves data from the handheld mobile system through the services from server. It applies various business rules to the on-premise internal handheld mobility data, sends the data to the server, and endows with administration with a web-based application.

10.4.9 Handheld Application

The handheld application comprises multiple screens. Users should log to the application by inputting username and password. The data authentication logic is part of the business data from the system. The applications are self-configurable; it downloads and installs new software via the server from the middleware (Figure 10.3).

10.5 SUMMARY

The force of mobility on industry is apparent. In growing numbers, business users are expected to handle serious tasks and decision making in real-time, no matter where they work from. An enterprise that endows with field service support is expected to have real-time transparency into inventory positions, enabling it to accept fresh deals on the fly with real-time changes from the field. These quick and well-organized operations are quickly becoming the important way to do business in order to maximize not only internal business requirements, but also external customer facing units.

The propagation of network connectivity, standards, and handy devices has made all of this possible. By adopting mobility, companies today are transforming the supply chain — from inventory, tracking to field offerings like scheduling, delivery, and navigation. In the case of financial sector enterprises, it is becoming necessary that at least some vital information and some dealings are available on the mobile devices, whether the users are the internal field users or the customers.