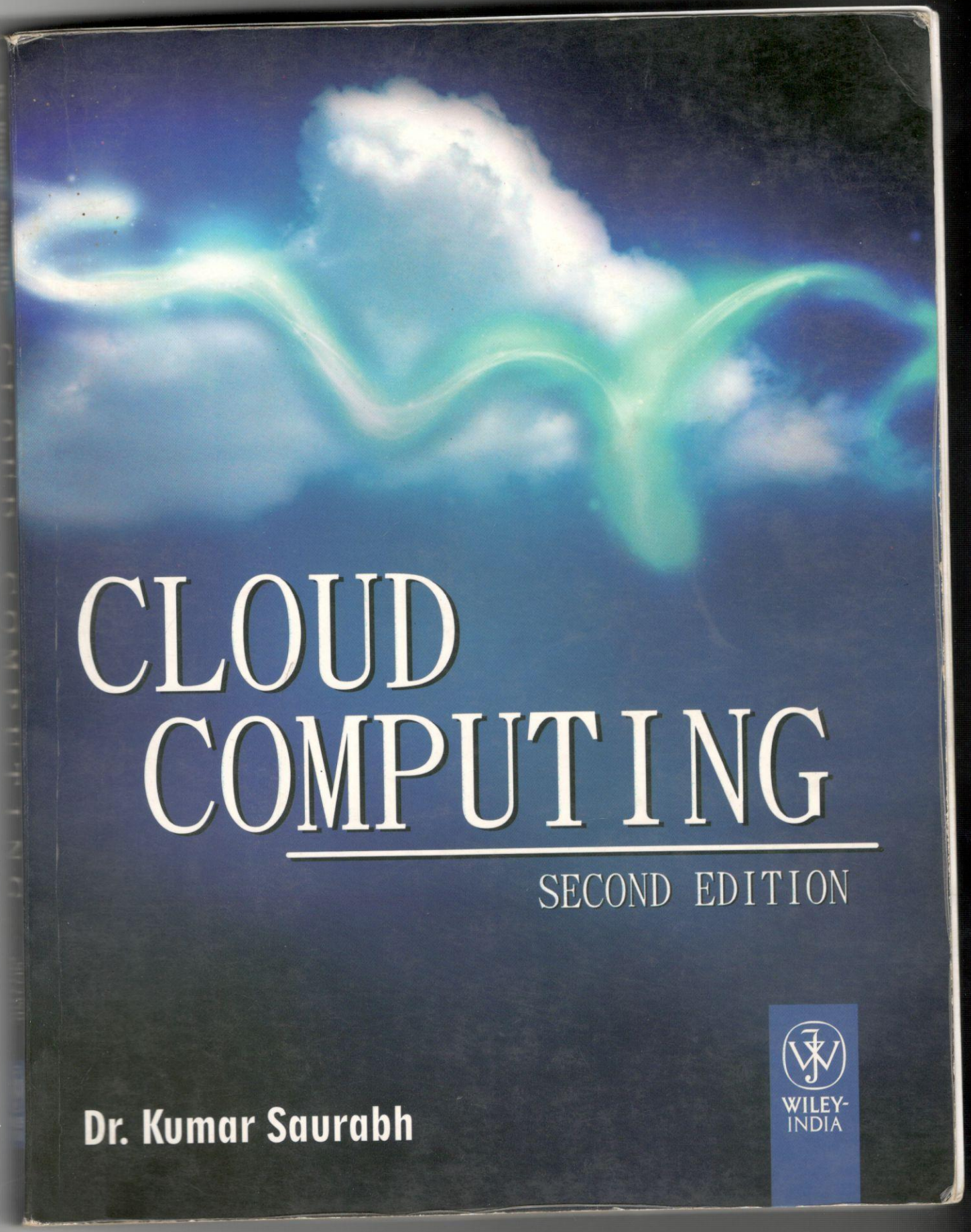

CLOUD
COMPUTING
SECOND EDITION
Dr. Kumar Saurabh
WILEY-
INDIA

Prologue

I got an opportunity to write the prologue for the first edition also and feel privileged to write a few words for the second edition as well. I compliment my dear colleague Dr. Kumar Saurabh for taking time from his packed schedules (at office) to document his recent real-time experiences, in this revised edition. This edition is fully revised to reflect the recent technological changes/trends.

Cloud Computing is becoming the term that describes the means of delivering any and all Information Technology components — from computing power to computing infrastructure, applications, business processes and collaboration – actually offering *IT as a Service*. Cloud computing is an emerging style of computing where applications, data and resources are provided to users as services over the Web. Technology is changing every day and organizations are expected to adopt the changes and transform enterprise IT — with self-service, charge back, service catalogs, physical resource orchestration, complete application provisioning, Hybrid IT, reservations, etc. It automates the self-service assembly and infrastructure management.

Dr. Kumar Saurabh has articulated his views through detailed examples of many aspects of cloud computing. He provides a new perspective to the field of cloud computing by discussing service and deployment models, its vitals, essence of virtualization, its penetration to SOA, mobility and infrastructure.

The chapters are significantly revised with newly added Cloud Architecture/Models and use cases. A new section on Cloud Analytics is added to prioritize automation of business processes, data analytics and business intelligence, as triggers to increase productivity and areas where investments are likely to be made in the near future. The section on SOA and cloud is revised to reflect SOA Governance throughout the *plan-build-run service* lifecycle, anchoring the process with strong policy governance.

Now, with the IT consumerization and adoption of mobile devices, consumers require services to be available globally, 24 × 7, 'on demand', massively scalable and 'pay as you grow' model. Consumers of a service would only care about what the service does for them, and not how it is implemented. This calls for a new paradigm; hence, the new chapter on mobility —to give a glimpse of real-time business environment across the value chain, by extending the boundaries of the enterprise data center to mobile devices.

This book is a treasure for system architects, solution architects, pre-sales managers, technical managers and infrastructure professionals, as it elucidates specific solutions to the real-world challenges. It helps the user community to understand the features and drawbacks of cloud computing through varied examples. It imparts the knowledge on cloud computing

and defines a new way to manage IT resources – enabling self-service provisioning of IT resources, metering-style accounting based on usage/time, automation of IT management in a standard process-driven environment. This would also serve as a ready-reckoner for students and faculty involved with IT curriculum.

I heartily compliment Dr. Kumar Saurabh for all his efforts and achievements and wish him all the very best for his future endeavors!

AS Murty
Chief Technology Officer
Mahindra Satyam

Preface

WHY THIS BOOK?

Cloud computing is a term that describes the means of delivering any and all Information Technology – from computing power to computing infrastructure, applications, business processes, and personal collaboration – to the end-users as a service wherever and whenever they require it.

When technology succeeds–when it meets the wishes of the people who use it, when it performs impeccably over a long interlude of time, when it is easy to adapt and even easier to employ–it can and does revolutionize things for superior reasons. But when technology fails–when its users are disgruntled, when it is error-prone, when it is complicated to transform and even difficult to use–dreadful things can and do happen. We all want to build technology that makes things healthier, avoiding the bad things that creep around in the shadow of disastrous hard work. To succeed, we need regulation when technology is architected, designed and built.

In the last 3 years since the first edition of this book was written, cloud computing has evolved from an incomprehensible idea practiced by a fairly low number of aficionado to a legitimate business computing discipline. Today, it is acknowledged as a subject commendable of serious research, reliable study, and turbulent debate. Although managers and practitioners identically identify the need for a more disciplined approach to cloud computing, they prolong to deliberate the manner it is to be applied to the business domain. Many individuals and companies are still adopting cloud technologies without proper cloud ROI and readiness assessments, even as they build systems to service today's most advanced technologies. Many professionals and students are unaware of modern methods of new era of infrastructure. And as a result, the quality of the architectures that we architect suffers, and terrible things happen.

This new edition book would be useful for developer, modeler, thinker, analyst, architect, researcher executive, manager, or any other IT professional already having experience of computing facts; or for those interested in virtualization, cloud offerings, and different types of real-world cloud implementation case studies. A developer making the transition from a conventional computing or traditional infrastructure to a cloud technology world; somebody already familiar with the general principles of cloud computing, but needs to know the drivers of cloud computing, its components, and its future for enterprise solutions; or someone who wants to make sure how cloud solution fits naturally into the real-world system and behaves as users expect them to all would find this new edition an apt resource.

A computing enthusiast should find in this book enough food for thought to start playing with the cloud solutions and concept implementations, and should be able to join the group of

technocrats continually working on new capabilities and performance enhancements. This book makes for a good introduction to cloud concepts and to cloud implementation in general.

Cloud computing is still a work in progress, and there's always place for new technocrats to jump into the game. This book is not intended to be a comprehensive guide or a reference to all aspects of cloud computing. Instead, we will be reviewing services, introducing the most important concepts and techniques, and giving examples of how to use them.

APPROACH

This book addresses the core issues of cloud computing, infrastructure, and virtualization. It presents everything one needs to know to be a successful system architect, technical manager, and infrastructure specialist for cloud computing. The book focuses on the real-world goals for services provided by cloud computing; the constraints on cloud computing infrastructure; the precise specification of cloud system structure and the implementation of specifications; the activities required in order to develop an assurance that the specifications and real-world goals have been met; and the evolution of cloud computing over time and across computing arena. It is also concerned with the processes, methods, and tools for the development of cloud infrastructure in an economic and timely manner.

Cloud computing, dynamic infrastructure, and virtualization have been deployed within every corporate function and within a broad section of businesses and markets. It involves changing paradigms about the way the world works, the way corporations function, and the human role in each case. The book reflects the core insights of cloud models, service offerings, and other benefits.

The widespread acceptance and deployment of cloud computing means virtualized cloud environments are now more critical than ever before. In the financial community as well as other market segments, even a relatively small system failure or outage can result in significant financial impact or have other far-reaching implications.

I have made essence-based revisions and transformed the organization in many of the chapters. Most importantly, we added one new chapter and reorganized the virtualization and cloud paradigms, model, standard and related concepts.

This new edition will give you the knowledge of important sections from the scratch, step-by-step procedures and the skills necessary to effectively understand the cloud computing concepts and implementation. It is meant to be very practical in nature and focuses only on the more important elements of cloud computing, not impenetrable subjects that have little relevance to the important issues faced by today's technocrats.

AUDIENCE

The second edition is intended to serve as a text to a maturing business era of cloud computing. The second edition, like the previous edition, is intended for both students and practitioners.

This book should be an interesting source of information both for people who want to experiment with cloud computing case studies and those who face the need to deal with the inne

levels of cloud infrastructure. We hope this book is useful as a starting point for people who want to experiment in the area of cloud computing, and review what cloud vendors are offering.

On the technical side, this text should offer a hands-on approach to understanding the cloud computing infrastructure, different types of real-world case studies and a techno-socioeconomic view of cloud infrastructure and some of the design choices made by the frontiers of cloud computing.

ORGANIZATION OF THE BOOK

In addition to many new and appreciably revised chapters, the chapters are ready with the new cloud computing paradigms, use cases and case studies for the ease of reading. All chapters are conceptualized with the relative success story based on standards, policies, compliances and benchmarks.

As I wrote this second edition, I was guided by the many observations and suggestions we received from readers of our previous edition, as well as by our own interpretation about the rapidly changing fields of cloud computing. The material in most of the chapters is well-run by bringing older fabric up-to-date and removing bits and pieces that was no longer of interest.

This revised book deals with the essentials of cloud computing. It visualizes the cloud benefits and services, and gives insights for cloud as a virtualization strategy. It discusses the set of hardware, software, networks, storage, services, and interfaces that combine to deliver aspects of computing as a service with shared resources, software, and information. This book works with concepts of cloud computing for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, software, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It will take care of different cloud models, layers, and different types of cloud offerings.

The new text focuses on cloud system management. It provides best practices to optimize the available resources and illustrates these practices by using practical examples. It also details how to secure and maintain virtual environments. It gives insights on the role of the virtualization technology in implementing the cloud infrastructure. It will aid in helping a client achieve the three *Cloud* delivery economic drivers: virtualization, standardization, and automation. The book also emphasizes on the future of cloud computing, the different cloud vendors, and cloud offering reviews. It discusses the business issues that provide compelling reasons regarding when to adopt the cloud offerings. It also gives the SOA-view of the enterprise cloud solution, its roadmap and driving forces. It describes how to monitor a virtualization infrastructure. Rather than presenting a list of tools, it addresses practical situations to help and use the monitoring tool that best shows the resources one is interested in. It also talks about the sizing aspects of the cloud infrastructure. The new edition is significantly more than a simple update. The book has been revised comprehensively and rationalized to emphasize new and important cloud computing processes and practices. In addition, a new chapter *"Cloud Mobility"* is added as companies are using mobile applications in greater than ever numbers to get better business dexterity and deliver greater customer values.

Kumar Saurabh

Acknowledgements

I take this opportunity to express my deep sense of gratitude to my research guide Dr V. B. Gupta, Devi Ahilya University, Indore, India for his enormous help, dynamic guidance and enthusiastic encouragement.

I am very much thankful to Mr. Vineet Nayyar, Honourable Chairman, Mahindra Satyam, for evincing keen interest in my work and continuous encouragement. I would like to thank Mahindra Satyam, especially Mr. C P Gurnani, Honourable Chief Executive Officer, Mahindra Satyam and Mr. A. S. Murty, Honourable Chief Technology Officer, Mahindra Satyam for motivating and encouraging me to work. I take immense pleasure in offering my acknowledgements to Mr. Sudhir Nair, Sr. Vice President, Infrastructure Management Services, Mahindra Satyam; Mr. Pravin Bolar, Vice President, Infrastructure Management Services; and Mr. Rishi Ranjan, Head, SI and solutions, Infrastructure Management Services, Mahindra Satyam for their valuable guidance that not only acted as a source of inspiration, but also encouraged me to discuss the problems critically and provided concrete suggestions without which it would have been impossible to bring this work in the present form.

I wish to place on records my sincere thanks to my wife Ms. Anju Varshney for her unconditional support and constant encouragement throughout this book. No form of acknowledgement can encompass the multitude of contribution that my family members have made. Their support, motivation, affection, and patience made me to work hard and complete this book successfully. Special thanks to my mother and father for their continued love and patience. I cannot forget the continuously support, affection and love provided by my younger brother Mr. Anuj Saxena.

I am very much grateful to my lovely daughters Shivika Saxena and Ashna Saxena who in their own ways provided me support and comfort.

Thanks are also due to all the colleagues at Mahindra Satyam and many others who directly or indirectly helped me in completing my book.

Finally, I thank Wiley-India for their continuing support, especially Paras Bansal (Publisher), Meenakshi Sehrawat (Senior Developmental Editor).

Dr. Kumar Saurabh
Bangalore
June, 2012

Special recognition towards the subject matter expertise brought in the book by

Mr. Rishi Ranjan

Assistant Vice President

Infrastructure Management Services, Mahindra Satyam

for authoring

Desktop Service: A VDI Perspective

About the Author

Dr. Kumar Saurabh has several years of industry experience with companies like Mahindra Satyam and IBM. Currently he is Sr. Consultant – Cloud Computing and Virtualization at Mahindra Satyam. He was Technical Consultant with IBM and worked on Sizing for IBM hardware platform and servers.

Saurabh is M. Tech and Ph.D. with specialization in System Dynamics Simulation and Modeling from Devi Ahilya University, Indore, India. He is also M. Sc. in Computer Science and has a PMP® certification. Author of more than 20+ research papers that have been published in various International, National journals and proceedings, presented in various international forums and he has also authored the book "*Unix Programming – The First Drive*" published by Wiley-India. Saurabh's research interests include virtualization, cloud computing and project management. His main area of research is System Dynamics and Simulation, Process Synchronization and Optimizations of UNIX Kernel and Device Drivers. Saurabh has also handled several large-scale systems performance and planning projects in a wide range of professional environments such as science and research, software companies and R&D products.

Contents

FIRST DRIVE

CHAPTER 1

Introduction

Essentials

Benefits

Why Cloud?

Business and IT Perspective

Cloud and Virtualization

Cloud Services Requirements

Cloud and Dynamic Infrastructure

Cloud Computing Characteristics

Cloud Adoption

Cloud Rudiments

Summary

1.1 INTRODUCTION

One of the latest drift in small and medium businesses and enterprise-sized IT is the need for a significant transformation of the IT environment. Cloud computing provides a major shift in the way companies see the IT infrastructure. This technology is primarily driven by the Internet and requires rapid provisioning, high scalability, and virtualized environments. It provides the abstraction for the business and is handled by the actual owners of the infrastructure experts. In this demanding world, the raison d'être to adopt cloud computing over standard IT deployments is flexibility, stability, rapid provisioning, reliability, scalability, and green solutions. Cloud computing can trace its intellectual roots back to grid computing, but it is often confused as the outcome of grid computing advancements and research during recent period, and that is not totally true. Grid computing paves the path for the evolution of the cloud computing concept. While these may be examples of applications of cloud computing for IT infrastructure, they are not the only ingredients of it. So, before going into the details of cloud computing, let us have a cursory glance at grid computing that gives you an immense computing grid to tap into as you need it, and scale up and down as per the requirement.

Grid computing approach starts with the breaking of the silos by inserting an additional layer on each server included in the grid. The main function of this additional layer is to create logical servers that distribute over different physical servers the computational needs (job, tasks) required by the different applications they are virtually executing. In this way, it is possible to decouple the applications from the physical systems on which they were running, and at the same time, it is possible to dynamically increase or decrease the computational power of the logical servers as per application needs.

1.1.1 Grid Computing

A grid is made up of a number of resources and layers with different levels of implementation (Figure 1.1). As said, there are different types of grid that are usually organized according to this taxonomy. Starting from the layer at the bottom – virtualization, which involves only physical resources – we may have then:

- **Information grids:** These are aimed to provide an efficient and simple access to data without worries about platforms, location, and performance.
- **Compute grids:** These exploit the processing power from a distributed collection of systems.
- **Services grids:** They provide scalability and reliability across different servers with the establishment of simulated instance of grid services.
- **A mix of them:** Each of these have specific sets of characteristics that are peculiar of the hybrid characteristics of compute and service grids.

Conceptually, we can imagine three layers, the lower being the physical one where we have the servers, storage devices, and the interconnecting network. In the second layer, we see the different operating systems, mapped one-to-one with the servers. The upper layer is the application one where we map different applications supporting the enterprise business processes.

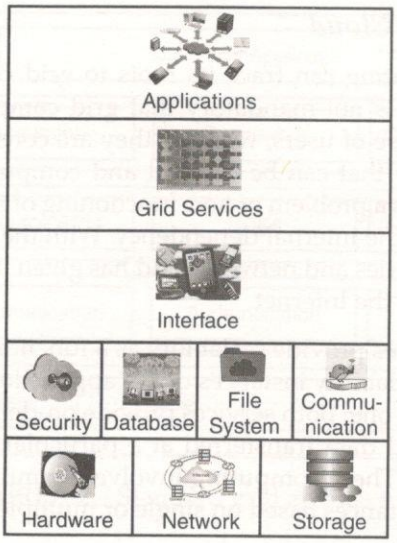


FIGURE 1.1 Simple grid architecture.

Grid computing is an evolution of distributed computing that utilizes open standards to allow you to see independent and physically scattered computing resources as though they were a unique large virtual computer.

With these concepts in mind, we can consider a 'compute grid', where the grid's goal is to exploit the processing power from a distributed collection of systems. Main functionalities of a compute grid are to manage the resources' workload, apply utilization policies and security rules, schedule and execute parallel tasks across distributed resources, and provision (reserving, adding, removing) resources according to the scheduling needs. It is a special kind of compute grid where resources – typically distributed all over the world, but could also be within an enterprise – are used by the grid only when idle, which means provisioning and scheduling policies are very 'relaxed'.

Information grid provides transparent and efficient access to data independent of their location, type, and platform, and allows end-users secure and transparent access to any information source regardless of where it exists. It supports sharing of data for processing and large-scale collaboration, and provides logical views of data without having to understand where the data is located or whether it is replicated. It manages data cache or data replication automatically to get the most efficient and secure access.

Information is usually defined as 'meaningful data' from the perspective of the end-user. An information grid provides an abstraction over disparate and distributed information sources, such as a Database Management System (DBMS), flat files (for example, comma-separated files), structured files (for example, XML documents), or a Content Management System (CMS).

An information grid also has the ability to federate or integrate data and information from heterogeneous resources into a unified virtual repository. The whole idea is to present a single view of the information.

1.1.2 Grid – The Way to Cloud

The concept of cloud computing can trace its roots to grid computing that provides rapid provisioning of resources. It is not mandatory that grid computing should be in the cloud; actually it depends on the type of users, whether they are consumers or administrators. Grid computing requires software that can be divided and computed or serviced on a single or multiple systems. This creates a problem of non-functioning of the overall solution if one of the components fails because of the internal dependency. With the advent of Internet, computing crossed geographical boundaries and networks and has given us the chance to exploit services and computing globally over the Internet.

Both cloud and grid services provide scalability as a functionality. This is achieved through load balancing and high availability instances of the applications running either on variety of operating systems or a single one. Both services provide on-demand services for the instances, users, storage, networks, and data transferred at a particular time, and can be de-allocated when they are not required. These computing involve the multi-tasking environments available on single or multiple instances based on single or multiple servers.

Optimization is a grid type where the primary focus is optimization of underutilized IT resources in an organization. Grids require a different way of thinking about how to deliver IT datacenter services, and resistance to changing behaviour is always the toughest hurdle to overcome in technology adoption. Lack of industry standards is a barrier to widespread adoption, as clients perceive the risk of not-protected technology investment. Security will have to be proven over time to potential customers at a number of levels for grids to be considered for adoption in shared workload environments. The cost of computational power (both CPU and storage) continues to decline, which may erode part of the financial benefits of grids. To exploit grid advantages fully, physical resources across heterogeneous systems can be virtualized building a single resource image.

The following sub-sections will help us understand the benefits of grid computing when deployed for infrastructure management and extended to cloud computing arena (Figure 1.2). This is also discussed in detail later in the chapter.

Storage/Data/Information

- Provides logical views of data without having to understand where the data is located or whether it is replicated.

System Management

- Defines, controls, configures, and removes components and/or services (could be physical) on a grid using automated or physical methods.

Metering, Billing, and SW Licensing

- Provides tools to monitor and distribute the number of licenses while using licensed software.
- Provides metering and billing techniques, such as utility-like services, so that the owners of the resources made available are accurately compensated for providing the resources.

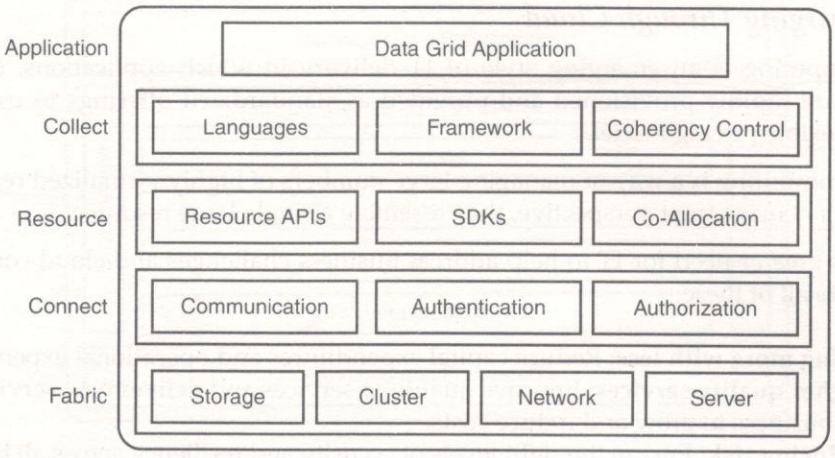


FIGURE 1.2 Standard grid architecture.

Security

- **Authentication:** The grid has to 'be aware' of the identity of the users who interact with it.
- **Authorization:** The grid has to restrict access to its resources to the users who are eligible to access it.
- **Integrity:** Data exchanged among grid nodes should not be subject to tampering.

Differing grid solutions may hit differing stages, but majority of the grid marketplace is transitioning from the 'early adoption' to the 'early majority' phase. Over the past few years, the market has evolved from specialist customers – predominantly in the academic and research sectors – using grid to accelerate internal simulations to a stage where corporate users are starting to apply grid and virtualization in a meaningful way that delivers clear business benefit (risk and portfolio analysis, seismic applications, clash analysis, etc.).

Organizations are now starting to use grid and virtualization technologies to unleash idle computing capacity to accelerate critical business processes and to optimize and improve resiliency of their IT infrastructure.

1.2 ESSENTIALS

Cloud computing is a term that describes the means of delivering any and all Information Technology – from computing power to computing infrastructure, applications, business processes and personal collaboration – to end-users as a service wherever and whenever they need it.

The *Cloud* in cloud computing is the set of hardware, software, networks, storage, services, and interfaces that combine to deliver aspects of computing as a service. Shared resources, software, and information are provided to computers and other devices on demand. It allows people to do things they want to do on a computer without the need for them to buy and build an IT infrastructure or to understand the underlying technology.

1.2.1 Emerging Through Cloud

Cloud computing is an emerging style of IT delivery in which applications, data, and IT resources are rapidly provisioned and provided as standardized offerings to users over the web in a flexible pricing model.

Cloud computing is a way of managing large numbers of highly virtualized resources such that, from a management perspective, they resemble a single large resource.

There is greater need for IT to help address business challenges and cloud computing can help you do all of these:

- **Doing more with less:** Reduce capital expenditures and operational expenses.
- **Higher quality services:** Improve quality of services and deliver new services that help the business to grow and reduce costs.
- **Reducing risk:** Ensure the right levels of security and resiliency across all business data and processes.
- **Breakthrough agility:** Increase ability to quickly deliver new services to capitalize on opportunities while containing costs and managing risk.

Cloud computing is the provision of dynamically scalable and often virtualized resources as a service over the Internet *(public cloud)* or intranet *(private cloud)*.

1.3 BENEFITS

As an emerging IT delivery model, cloud computing can significantly reduce IT costs and complexities. The buzz surrounding cloud is based mostly on a new kind of user experience – particularly in the consumer Web space – for search, social networking, and retail. From the consumer perspective, cloud computing is a means of acquiring services without needing to understand the underlying technology. Many of us use cloud delivery models everyday without knowing it when we share photos online, download music, or access bank accounts using our mobile phone.

From a technology perspective, cloud computing is loosely defined as a style of computing where dynamically scalable resources (such as CPU, storage, or bandwidth) are provided as a service over the Internet. The process is typically automated and takes minutes. Cloud computing can be considered as a massively scalable, self-service delivery model that lets you access processing, storage, networking and applications as services over the Internet. Enterprises adopt cloud models to improve employee productivity, deploy new products and services faster and reduce operating costs – starting with workloads that are ripe for this environment. These typically include development and test, virtual desktop, collaboration, and analytics.

1.4 WHY CLOUD?

A cloud typically contains a significant pool of resources, which could be reallocated to different purposes within short time frames, and allows the cloud owner to benefit significantly from

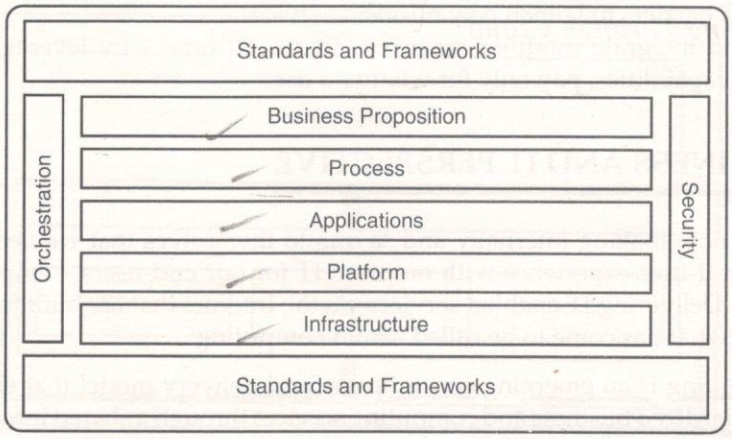


FIGURE 1.3 Basic cloud computing model.

economies of scale as well as from statistical multiplexing (Figure 1.3). The entire process of requesting and receiving resources is typically automated, and is completed in minutes.

Cloud services today are delivered in a user-friendly manner and offered on an unprecedented scale. The payment model is pay-as-you-go and pay-for-what-you-use, eliminating the need for an up-front investment or a long-term contract. This presents a less disruptive business opportunity for businesses with spiky or unpredictable IT demands, as they are able to easily provision massive amounts of resources on a moment's notice and release them back into the cloud just as quickly.

There are different reasons for adopting the cloud:

- Massive, Web-scale abstracted infrastructure.
- Dynamic allocation, scaling, movement of applications.
- Pay per use.
- No long-term commitments.
- OS, application architecture independent.
- No hardware or software to install.

This results in business- and IT-aligned benefits:

- Accelerate innovation projects that can lead to new revenue.
- Make IT an enabler of, not a barrier to, rapid innovation.
- Provide an effective and creative service delivery model.
- Deliver services in a less costly and higher quality business model, while providing service access ubiquity.
- Create a sustainable competitive differentiation.
- Rapidly deploy applications over the Internet and leverage new technologies to deliver services when, where, and how your clients want them – before your competitors do.

- Lower IT barriers to launch new business services.
- Build and integrate modular services – in record time – by leveraging 'rentable' IT services capabilities, pay only for what you use.

1.5 BUSINESS AND IT PERSPECTIVE

Businesses are now looking internally and saying to themselves that we need to deliver this same level of end-user experience with our own IT for our end-users – employees, partners, and customers. Delivering IT-enabled services via the Internet that are built for the end-user to be in control is what has come to be called 'cloud computing'.

Cloud computing is an emerging consumption and delivery model that enables the provisioning of standardized business and computing services through a shared infrastructure, where the end-user is enabled to control the interaction in order to accomplish the business task.

Computing resources such as processing power, storage, databases, and messaging are no longer confined within the four walls of the enterprise. Instead, a tightly woven fabric of abstract – or virtual – resources are tapped into whenever they are needed. Essentially, everything needed from a computing resources standpoint is provisioned by the cloud – much like the electrical power grid we all tap into.

1.6 CLOUD AND VIRTUALIZATION

Virtualization has been around for 30 years. Yet, how many have really truly virtualized at all the layers of the stack? You really cannot expect cloud to produce what a cloud is expected to produce if it is not virtualized, standardized, and automated, because people expect scalable services.

In a cloud environment, people expect self-service, being able to get started very quickly, self-provisioning, or rapid provisioning. All of those things essentially demand that you do have these very important fundamentals in place.

The only way you are going to be able to get efficiency is by virtualizing, standardizing, and automating (Figure 1.4). And that's going to drive down costs and improve service. This is really a pretty simple equation and we are seeing organizations that are doing this achieve very real measurable business results. These results include:

- **Server/storage:**
  - IT resources from servers to storage, network, and applications are pooled and virtualized to help provide an implementation-independent, efficient infrastructure, with elastic scaling – environments that can scale up and down by large factors as demand changes.
- **Automation using:**
  - Self-service portal: Point and click access to IT resources.
  - Automated provisioning: Resources are provisioned on demand, helping to reduce IT resource setup and configuration cycle times.

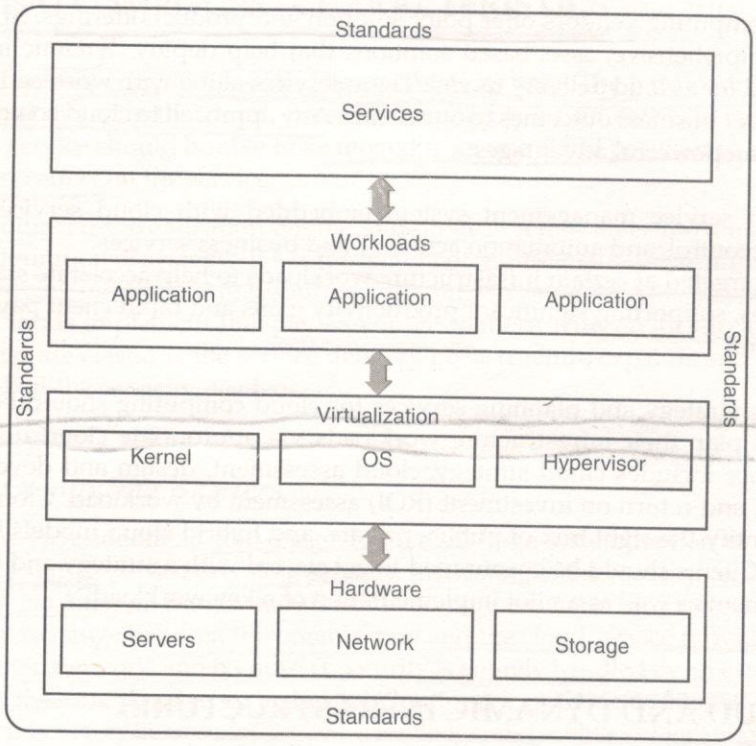


FIGURE 1.4 Datacenter clouds.

- **Standardization through:**
  - Service catalogue ordering: Uniform offerings are readily available from a services catalogue on a metered basis.
  - Flexible pricing: Utility pricing, variable payments, pay-by-consumption with metering and subscription models help make pricing of IT services more flexible.

1.7 CLOUD SERVICES REQUIREMENTS

Cloud computing is being touted as the next best thing for cutting the cost of providing first-class IT services. You can decide which workloads are right for the cloud and which may not be through an examination of your workloads – uses of IT resources for particular activities or tasks. You can also decide which workloads can go on the vendor cloud (via the Internet or a virtual private network [VPN]) and which need to remain onsite (behind the organization's firewall). This focus on outcomes and delivery models presents a new opportunity to open up competitive accounts and expand the IT optimization conversation with existing clients.

Most cloud computing vendors offer point-solution and product offerings. In contrast, one should offer comprehensive, asset-based solutions that help deploy dynamic infrastructure, which is required for a cloud delivery model. These services along with workload solutions are designed to deliver business outcomes to our clients. Any approach to cloud computing should offer the following powerful advantages:

- A proven service management system embedded with cloud services to provide visibility, control, and automation across IT and business services.
- Services targeted at certain infrastructure workloads to help accelerate standardization of services, supporting significant productivity gains and rapid client payback on their investment.

Infrastructure strategy and planning services for cloud computing should be designed to help companies plan their infrastructure workloads via appropriate cloud delivery model. Specific assistance includes cloud strategy, cloud assessment, design and development of a cloud roadmap, and return on investment (ROI) assessment by workload. Cloud leaders can help clients identify the right mix of public, private, and hybrid cloud models for infrastructure workload. Clients should be encouraged to get started with a strategy and planning consulting engagement as well as a pilot implementation of a key workload.

1.8 CLOUD AND DYNAMIC INFRASTRUCTURE

Through cloud computing, clients can access standardized IT resources to deploy new applications, services, or computing resources rapidly without re-engineering their entire infrastructure, thus making it *dynamic*.

Cloud Dynamic Infrastructure is based on an architecture that combines the following initiatives:

- **Service Management:** Provide visibility, control, and automation across all the business and IT assets to deliver higher value services.
- **Asset Management:** Maximize the value of critical business and IT assets over their lifecycle with industry tailored asset management solutions.
- **Virtualization and Consolidation:** Reduce operating costs, improve responsiveness, and utilize resources more fully.
- **Information Infrastructure:** Help businesses achieve information compliance, availability, retention, and security objectives.
- **Energy Efficiency:** Address energy, environment, and sustainability challenges and opportunities across the business and IT infrastructure.
- **Security:** Provide end-to-end industry customized governance, risk management, and compliance for businesses.
- **Resilience:** Maintain continuous business and IT operations while rapidly adapting and responding to risks and opportunities.

1.9 CLOUD COMPUTING CHARACTERISTICS

Cloud computing uses commodity-based hardware as its base. The hardware can be replaced any time without affecting the cloud. It uses a commodity-based software container system. For example, a service should be able to be moved from one cloud provider to any other cloud provider with no effect on the service.

This also requires a virtualization engine and an abstraction layer for the hardware, software, and configuration of systems. It has the feature of multi-tenant where multiple customers share the underlying infrastructure resources, without compromising the privacy and security of their data. Clouds implement the 'pay-as-you-go' pattern with no lock-in and no up-front commitment and are elastic as the service delivery infrastructure expands and contracts automatically based on the capacity needed.

1.9.1 Cloud Computing Barriers

IT organizations have identified four major barriers to large-scale adoption of cloud services. The first one is security, particularly data security. Interestingly, the security concerns in a cloud environment are no different from a traditional datacenter and network. However, since most of the information exchange between the organization and the cloud service provider is done over the web or a shared network, and because IT security is entirely handled by an external entity, the overall security risks are *perceived* as higher for cloud services. Some additional factors cited as contributing to this perception are (a) limited knowledge of the physical location of stored data, (b) a belief that multi-tenant platforms are inherently less secure than single-tenant platforms, (c) use of virtualization as the underlying technology, where virtualization is seen as relatively new technology, and (d) limited capabilities for monitoring access to applications hosted in the cloud.

The next one is *governance and regulatory compliance*. Large enterprises are still trying to sort out the appropriate data governance model for cloud services, and ensuring data privacy. Quality of service (availability, reliability, and performance) is still cited as a major concern for large organizations. Not all cloud service providers have well-defined service-level agreements (SLAs), or SLAs that meet stricter corporate standards. Recovery times may be stated as 'as soon as possible' rather than a guaranteed number of hours. Corrective measures specified in the cloud provider's SLAs are often fairly minimal and do not cover the potential consequent losses to the customer's business in the event of an outage. Inability to influence the SLA contracts is another issue. From the cloud service provider's point of view, it is impractical to tailor individual SLAs for every customer they support. The risk of poor performance is perceived higher for a complex cloud-delivered application than for a relatively simpler on-site service delivery model. Overall performance of a cloud service is dependent on the performance of components outside the direct control of both the customer and the cloud service provider, such as the network connection.

Integration and interoperability is the third concern. Identifying and migrating appropriate applications to the cloud is made complicated by the interdependencies typically associated with business applications. Integration and interoperability issues include a lack of standard interfaces

or APIs for integrating legacy applications with cloud services. This is worse if services from multiple vendors are involved. It also includes software dependencies that must also reside in the cloud for performance reasons, but which may not be ready for licensing on the cloud. There are worries about how disparate applications on multiple platforms, deployed in geographically dispersed locations, can interact flawlessly and can provide the expected levels of service.

The next concern is whether the workloads are suitable (or not) for cloud deployment. Not every application is a suitable candidate for moving to a cloud computing environment. Whether or not a particular application is a good fit depends on a combination of the nature of the business functions it provides, the capacity characteristics it requires (some processing patterns will be more cost-effective than others in a pay-as-you-use model), and technical aspects of the application or its infrastructure requirements.

1.10 CLOUD ADOPTION

Business function that suits cloud deployment can be low-priority business applications, for example, business intelligence against very large databases, partner-facing project sites, and other low-priority services. Cloud favours traditional Web applications and interactive applications that comprise two or more data sources and services and services with low availability requirements and short life spans; for example, enterprise marketing campaigns need quick delivery of a promotion that can just as quickly be switched off. It is also helpful when high volume, low cost analytics and disaster recovery scenarios, business continuity, backup/recovery-based implementation are required. It is like a boon to one-time batch processing with limited security requirements, record retention, media distribution, and mature packaged offerings, like e-mail, collaboration infrastructure, collaborative business networks.

Based on technical characteristics, we can say that it is suitable for applications that are modular and loosely coupled; isolated workloads; single virtual appliance workloads and software development and testing; and pre-production systems. It gels well with R&D projects, prototyping to test new services, applications, and design models and applications that scale horizontally on small servers – that is, by adding more servers, rather than by increasing a server's computational capacity.

Applications that need significantly different levels of infrastructure throughout the day, such those used almost solely during the business day, should be deployed through cloud. Applications that need significantly different levels of infrastructure throughout the month, or that have seasonal demand, such as those used primarily during the end-of-the-quarter close or during a holiday shopping season, are the best examples of cloud deployments. Applications where demand is unknown in advance – for example, a Web start-up will need to support a spike in demand when it becomes popular, followed potentially by a reduction once some of the visitors turn away – can also be deployed using clouds.

It is not suitable for mission-critical and core business applications, transaction processing and applications that depend on sensitive data normally restricted to the organization, or requiring a high level of auditability and accountability as these process cannot share the high importance data, processing power, and hardware with the third party. Applications that run 24×7×365 with steady demand, applications that consume significant amounts of memory,

including applications dependent on large in-memory caches, databases, or data sets are not suitable for cloud. Applications that take full advantage of multiple cores, such as those that do a significant amount of parallel processing, and thus benefit from many cores on a single server, are not recommended for cloud deployment.

It is not recommended for applications that require high-performance file system I/O needing high-bandwidth interserver communications, for example, highly distributed applications. Cloud does not work well with applications that scale vertically on single servers – that is, by increasing a server's computational capacity rather than adding more servers and applications dependent on third-party software, which does not have a virtualization or cloud aware licensing strategy.

1.11 CLOUD RUDIMENTS

Cloud delivers a software platform that will enable customer IT to build an Infrastructure-as-a-Service (IaaS) cloud. Cloud is built on the capabilities of existing virtualization management and physical server provisioning solutions to deliver application infrastructure to users that can be consumed in a self-service manner.

Cloud optimizes the usage of the physical and virtual infrastructure through intelligent resource allocation policies, and adds the ability to flex applications elastically based on demand. The high-level capabilities of any cloud include the following:

- **Resource Aggregation and Integration:** Cloud solution operates on top of existing virtualization management, physical server provisioning, and system management environments. It retrieves inventory information about machines and software templates from multiple locations, and aggregates this information into a central logical view of all resources in the infrastructure.
- **Application Services:** Rather than provide access to resources directly, cloud solutions' application 'Definitions' describes packages of machine capacity and software images that can be allocated by resource consumers. Applications can range from individual machines provisioned with an operating system image through to full multi-tier application environments that consist of collections of machines and software stacks provisioned in a specific order with network and storage dependencies handled through integration with third-party management tools. Application instances represent an agreement between the cloud provider and consumer to use capacity on a reservation or on-demand basis. Reservations allocate capacity in the resource inventory, guaranteeing that the capacity will be available to the consumer at some defined point in the future. On-demand allocations provide access to resources but do not guarantee availability. Reserved and on-demand capacity can be combined in an application, where a baseline of capacity can be elastically increased or decreased according to metrics and policies defined by the consumer.
- **Self-Service Portal:** An important principle of a cloud solution is to enable self-service access to resources with minimal IT involvement. It should support the notion of account owners signing up for contracts and then being able to delegate the use of the purchased capacity within their own groups or departments. Users can request machines

or entire multi-machine application environments and monitor and control them using a web-based self-service portal. The system will drive the workflows necessary to create the environment, and provide run-time environment management in order to support application elasticity.

- **Allocation Engine:** Dynamic Resource Management (DRM) is the automated allocation and reallocation of IT resources based on policies that express business demands and priorities. DRM is a key component of any cloud solution that maximizes the efficiency of the IaaS infrastructure. DRM policies should be applied both when initially placing applications onto machine resources and when selecting applications to migrate in order to preserve SLAs around application performance. Some of the allocation and migration strategies include advance reservation of resources, load-based placement and migration, application and resource topology constraints, energy usage optimization, etc. The use of sophisticated DRM helps to increase utilization of cloud resources, reduces overspending by effectively using existing resources, and saves costs in terms of operations, power, and cooling.
- **Reporting and Accounting:** In order to close the loop and determine how the cloud is behaving, metering information on resource allocations as well as actual usage is collected in an accounting database. The data is centrally available to create reports on inventory capacity, capacity allocated versus capacity used by contract, and usage-billing reports based on consumed resources.

The following are the *cloud features* that would help to bring in *agility* and *transparency* along with increase in the utilization of the existing resources at the datacenter of any customer.

- **Self-Service:** This feature presents an interface for separate authenticated end-users – via role-based access controls (RBAC) – to select options for deployment. It should have unique policy controls per tenant and user role, and the ability to present unique catalogues per user or group. The self-service portal is a web interface also accessible in other ways, such as through a mobile client, etc.
- **Dynamic Workload Management:** With cloud solution implemented, datacenters are enabled with automation and orchestration software that coordinates workflow requests from the service catalogue or self-service portal for provisioning virtual machines. Also each provisioned virtual machine is enabled with a life-cycle for deployment expiration which increases the efficiency of utilization of resources.
- **Resource Automation:** Using cloud solution, Admins or engineering team members of the datacenter could control the heterogeneous environment on a single pane. This feature establishes secure multi-tenancy, isolates virtual resources, and helps prevent contention in the load aware resource engine which intelligently does the workload packing or load balancing across hypervisors automatically.
- **Chargeback, Showback, and Metering:** Using this feature Admins could bring out the usage reports for cloud infrastructure service consumption and these usage reports serves as a basis for metering and billing system. Using this Admins will be able to understand if the virtual machines are attached with appropriate resources. Enabling *chargeback, showback,* and *metering* in any organization would bring in transparency to the business and environment for management to clearly see the usage and dollar value associated to it and take decision-making steps.

- **Open Architecture:** The cloud should be integrated with existing third-party products that are already installed in the datacenter. It should also be integrated to a public cloud for using additional resources and should be managed through a single cockpit. It is also possible to meter the public cloud resource usage.
- **Image Pools:** The cloud solution should have full blown service catalogue and support to most of the operating systems. It should be possible to vary the hardware configuration for the templates. It should also integrate with existing templates and images used by the development and testing teams.
- **Role-Based Access Administration:** The cloud solution should have the capability to integrate cleanly with any of the existing, LDAP, or other authentication and identity mechanisms. These features are crucial for providing secure multi-tenancy. This would also bring in security to the self-service portal.
- **Virtualization:** The cloud should extend support to virtualization layer. This implies that it should support most of the industry-proven hypervisors. This enables the Admins and engineering team of the datacenter to control them over a single pane.

1.11.1 Cost Savings with Cloud

Faster Time-to-Market (Missed Business Opportunity)

Deploying new application environments quickly and reliably can have a directly impact on competitiveness enabling organizations to take market share. The cloud will enable automated delivery of application environments exponentially faster than current practices.

With the cloud model, teams could be delivering fully configured, multi-component application environments to users in some minutes. This makes an immediate impact on user efficiency as well as eliminating much of the manual labor previously required of both the IT and application teams. In addition to this, ability to remove a (physical or virtual) will have similar performance, and once again allows that compute power to be available for other uses.

Public Cloud Interfaces

Cloud infrastructure with its policies should manage workload placement optimally by looking at several metrics. Cloud should also offer the capability to burst out to public cloud or internal resources when needed and cut off that link when done. The cloud should also be able to meter for the usage of the deployed instances in public cloud. Customer datacenter could use resources in public cloud for test and development environment if there are no resources available on the premise which will also help them to defer from the new hardware procurement.

Automated Scaling

The cloud solution should provide an out-of-box functionality to flex-up or flex-down an application instance or resource based on performance metrics and should also flex-up and flex-down an environment automatically or manually. The cloud solution should offer policies that can be customized to look at any metric and take action based on the threshold. These policies

must be embedded in a service catalogue to monitor an application or the entire environment and flex-up or flex-down with more resources.

Business Transparency

Service Accounting helps to improve utilization of datacenter infrastructure with accurate visibility into the true costs of physical and virtualized workloads. It will enable decision makers to have full cost transparency and accountability for usage, metrics, roles and definitions. This would also help an Admin to understand whether a machine is equipped with right resources or not.

1.11.2 Benefits

Cloud brings lot of benefits for any enterprises. Let us explore these in brief here. They will be discussed in detail in the later chapter also.

- Increase agility on the IT datacenter resources and innovation.
- Enable self-service portal and thus ensure VM in less lead-times.
- SLAs are met as the VM lead-times and downtimes are significantly reduced.
- Trial and error configuration tests can be done at ease.
- Complete control over cloud usage for Admins.
- Scalability and flexibility allow the IaaS cloud to almost deliver the promise of unlimited IT services on demand.
- Pay for only what they use and are not charged when their service demands decrease.
- Significant reduction in the costs for IT datacenter.
- Private cloud enables dynamic sharing of the resources available in IT datacenter so that demands can be met cost-effectively.
- Considerable increase in the utilization of resources of IT datacenter.
- Increase in operational efficiency of the resources in the IT datacenter.
- Achieve a greener datacenter (server consolidation and virtualization enables over committed machines).
- Support for heterogeneous hardware vendors. Avoids Vendor Locking.

It will help the enterprises by

- Reducing the number of administrators required to manage a more diverse IT resource pool.
- Dramatic reduction in cycle times to provision new assets.
- Realization of an infrastructure 'pay-per-use' model.
- Reduction in planned capital spending and maintenance.
- Increased user satisfaction with IT services.
- Reduction in physical server count.
- Consolidation of enterprise application licenses.
- Flexibility to meet future demands on infrastructure goals that can be leveraged.
- Capacity on-demand (pre-provision, automate).

- Consolidated, streamline change control.
- End-to-end application provisioning.
- Allowing developers to provision development application environment autonomously.
- End-to-end performance measurement.
- Consumption-based charge back.
- Plan for active/active datacenter operations.
- Plan for increased datacenter density.
- Separate production and development networks.

1.12 SUMMARY

In this chapter, we explored cloud computing, its benefits, and its services. The chapter also gave deep insights into cloud computing models that are put into practice. The next chapter discusses the different types of cloud models and service platforms.

CHAPTER 2

CLOUD DEPLOYMENT MODELS

2.1 INTRODUCTION

Cloud computing is an emerging style of computing where applications, data, and resources are provided to users as services over the Web. The services provided may be available globally, always on, low in cost, 'on demand', massively scalable, 'pay-as-you-grow'. Consumers of a service need to care only about what the service does for them, and not on how it is implemented. Cloud computing is a technology that allows users to access software applications, store information, develop and test new software, create virtual servers, draw on disparate IT resources, and more – all over the Internet (or other broad network).

Cloud computing is a model-driven methodology that provides configurable computing resources such as servers, networks, storage, and applications as and when required with minimum efforts over the Internet services. Cloud also indicates essential characteristics, delivery models, and deployment models.

This chapter visualizes several models for cloud computing, including private clouds (where the deployment is within the organization's firewall) and public clouds (where the application services and data are hosted by a third party outside the firewall). Consistent data availability and security is a critical success factor for any cloud deployment. Businesses need to ensure that data is adequately protected and can be restored in a timely fashion following any disruption event.

Clouds need a datacenter, but the aim of cloud computing is to eliminate the need to think about datacenters. A datacenter is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression), and security devices.

Datacenters are tied to locality, with specific components including redundant power supplies, redundant communications, environmental controls, security devices, etc.

Clouds are location-independent, providing abstracted versions of datacenter components that are not tied to a specific datacenter: virtual servers, virtual storage, virtual networking, etc. Reliability and redundancy comes from cloud providers using multiple datacenters, so clouds almost certainly span one or more datacenters, but themselves are not datacenters.

2.2 CLOUD CHARACTERISTICS

Cloud carries the basic infrastructure characteristics that are helpful to deploy cloud service in a fast and cost-effective way (Figure 2.1). The following characteristics set apart cloud from other computing techniques.

2.2.1 *On-Demand Service*

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

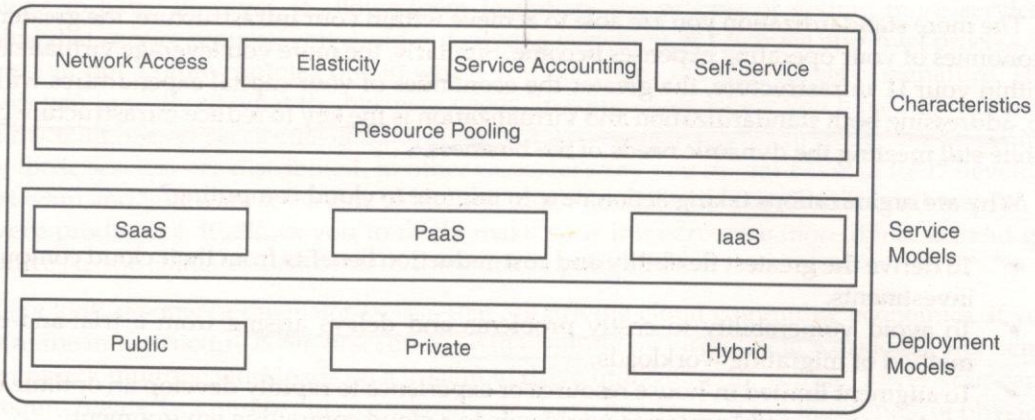


FIGURE 2.1 Cloud model.

2.2.2 *Ubiquitous Network Access*

Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops and personal digital assistants [PDAs]).

2.2.3 *Location-Independent Resource Pooling (Multi-Tenant)*

The provider's computing resources are pooled to serve multiple customers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to the demand. There is a sense of location-independence in that the customer generally has no control or knowledge about the location where the services are located (for example, country, state, or datacenter). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

2.2.4 *Rapid Elasticity*

Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

2.3 MEASURED SERVICE

Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and the consumer of the utilized service.

The more standardization you are able to achieve within your infrastructure, the greater the economies of your operating expenses become. Similarly, the more you leverage virtualization within your IT infrastructure, the greater the economies of your capital expenditures will be. So, addressing both standardization and virtualization is the key to reduce infrastructure costs while still meeting the dynamic needs of the business.

Why are organizations taking action now to migrate to cloud computing?

- To derive the greatest flexibility and cost-reduction benefits from their cloud computing investments.
- To avoid vulnerability to costly problems and delays arising from a trial-and-error method of migrating workloads.
- To augment limited in-house resource or experience to rapidly develop an optimization roadmap to smoothly migrate workloads to a cloud computing environment.

Cloud vendors can address client's challenges by:

- Prioritizing workloads for cloud adoption based on business impact and risk.
- Maximizing business return by identifying applications that are well suited for cloud computing and have high business impact.
- Addressing problematic workloads to improve their propensity for cloud computing.
- Helping avoid costly implementation issues by identifying and addressing potential difficulties during migration.
- Mitigating the risk of costly implementation delays by identifying potential problems and addressing them before migration.
- Avoiding inadequate performance of highly complex and integrated workloads.
- Leveraging expertise to deliver an actionable roadmap to successfully migrate applications to a cloud computing environment.
- Accelerating your cloud initiatives.

2.3.1 *Cost Factor*

There are a number of reasons why cloud computing is popular with businesses. There is the cost aspect. By virtualizing your environment and standardizing it, you can deliver more services with fewer resources and drive up utilization. By adding automation, you can reduce labour cost giving you an additional cost benefit. This gives you a lot of flexibility because you can access cloud workloads services without thinking about the location and time of its execution. What this allows the organization to do then is to free up budget so that the money can be diverted to innovation and development of new capabilities rather than just keeping the lights on and running the IT enterprise.

The growing complexity of IT systems and soon a trillion connected things demand that sprawling processes become standardized services that are efficient, secure, and easy to access. A service management system will provide visibility, control, and automation across IT and business services to ensure consistent delivery. Self-service plus standardization will drive lower operational costs, unlock productivity, and ensure better security.

Cloud allows businesses to be smarter about how they deliver services. The first aspect of this is a self-service portal. This allows your end consumers to only see services they are

Self service portal

allowed to have; however, it allows them to initiate the process of getting those services. Behind that service request, you could put either a very light or no-touch approval process, or you could put a more complex one in which you may need multiple levels of signature. This allows you to really fit what the business needs. In some cases where you have high security, you have high-level service-level agreements (SLAs) – you really want to be able to control how those services are distributed. In other cases, let's say you do not have an R&D development team and you want to be able to have as much flexibility as possible. This allows you to be very productive. It allows you to really make your infrastructure more dynamic, and get resources to the teams that really need them at any point in time.

Let's look at some of the major factors that are driving cloud computing economics. If you look at the infrastructure layer, first comes virtualization. By virtualizing workloads and being able to stack multiple workloads on a system to drive utilization up, you can lower your capital requirements. In a number of cases, businesses have hundreds, if not thousands, of physical servers and unless they have used virtualization and unless they are really driving that utilization, the utilization could be as low as 10 percent. So, in a lot of cases, organizations that use cloud computing are able to drive utilization up and either lower future capital requirements or even retire antiquated equipment and drive their costs down.

From a labour perspective, using a self-service portal allows your clients to help themselves. So there is less support and it makes the offering more available from a service perspective. In terms of automation, it takes tasks that are very manual and repeatable, and by automation then, it reduces your IT operations cost. In a development or test environment, you need multiple skills to get that environment to the end-user. You need operating system skills, middleware skills, database skills, and application skills. This allows you to define that environment as a repeatable, deployable resource, and it drives down your labour cost there. Of course, you need to standardize those workloads. Standardization has labour cost and quality benefits so that you can ensure consistency from environment to environment.

In many cases, you may want to use multiple models for different types of services that you want to deliver. Starting with private cloud services, the first model (which is also the most popular currently) is the private on-premise cloud. If the cloud is within the organization's datacenter, it is operated and managed by the organization itself.

The need to achieving cost optimization has also provided fertile ground for cloud computing. The cloud paradigm is an attempt to improve service delivery by applying engineering discipline and economies of scale in an Internet-inspired architecture.

Cloud computing can be an important new option in helping businesses optimize the IT expense equation while maintaining fast, high-quality service delivery.

2.3.2 *Benefits*

We can enjoy many benefits by adopting cloud:

- **Self-service capability:** Once somebody deploys the cloud services, they are capable of self-service. Now testing teams do not have to buy computing services as they can enjoy the same services over the cloud and it reduces the procurement process. Hence, they can concentrate on the testing services and efforts.

- **Resource availability:** It is the one of the most common benefit facilitated by virtualization. It also helps to track and leverage the resource pool under the same umbrella of resource units.
- **Operational efficiency:** Sometimes conventions and configurations followed by test and operation teams may differ from those followed by development teams. This can cause the application behaviour to be different from what was intended as well as delay services. The template-based approach, with its solution stacks of hardware, configurable applications, and operating systems, is more transparent and can help the teams to understand the environment better.
- **Hosted tools:** Due to these, the developers and testers need not install, configure, run, or maintain tools on their systems as they can log into the tools from any machine on the network maintaining the tools. Rather they can simply login to the tools and enjoy the services over the network.

These four benefits help the developers and testers to concentrate on their core work, retain focus, and concentrate more on their work without worrying about other jobs. This increases quality and productivity, and therefore, more developer innovation, increased test quality and coverage, etc. which are beneficial for an organization.

There are a number of major challenges developers face today in getting started and rolling out new applications and services faster. However, innovative new products and services are the lifeblood of rapidly growing companies. They represent a substantial portion of corporate sales and profits. In an environment of heightened competition, the inability to roll out new applications and services quickly means declining market share and lost revenue.

A growing application backlog leaves lines of business and end-users frustrated because they feel IT is a bottleneck and they look for ways to work around IT to roll out new products and services more quickly. Testing backlog is often very long, and a major factor in the delay of new application deployments.

A major reason testing takes so long is it takes weeks, on average, to set up application environments for test and QA as well as production. This is because of the time it takes to procure new hardware and software, and then schedule time with IT to configure and set up the systems. Configuration and setup are manual processes where errors are easily introduced. The average new application takes six to nine months to deploy, on average. This is caused by a number of factors ranging from poor governance to poor collaboration between business users and development to inflexible infrastructure and tools. Almost 30 percent of all defects are caused by wrongly configured test environments. This is a result of manual processes without any automation to replicate testing environment along with challenges organizations face in finding available resources to perform tests in order to move new applications into production. Test environments are seen as expensive and provide little real business value.

2.4 CLOUD DEPLOYMENT MODELS

Let's talk about cloud computing and the different types of cloud deployment models and different types of services that can be delivered using that model. Cloud computing is a style

of computing in which business processes, application, data, and any type of IT resource can be provided as a service to users.

Cloud delivery models can be briefly classified into three types (Figure 2.2):

- **Public:** In a public cloud, a business rents the capability and they pay for what they use on-demand.
- **Private:** In private clouds, a business essentially turns its IT environment into a cloud and uses it to deliver services to their users.
- **Hybrid:** Hybrid clouds combine elements of public and private clouds.

A private cloud drives efficiency while retaining control and greater customization. Public clouds today are for processes deemed more easily standardized and a lower security risk. There some functions that already exhibit a high degree of standardization, that are more easily

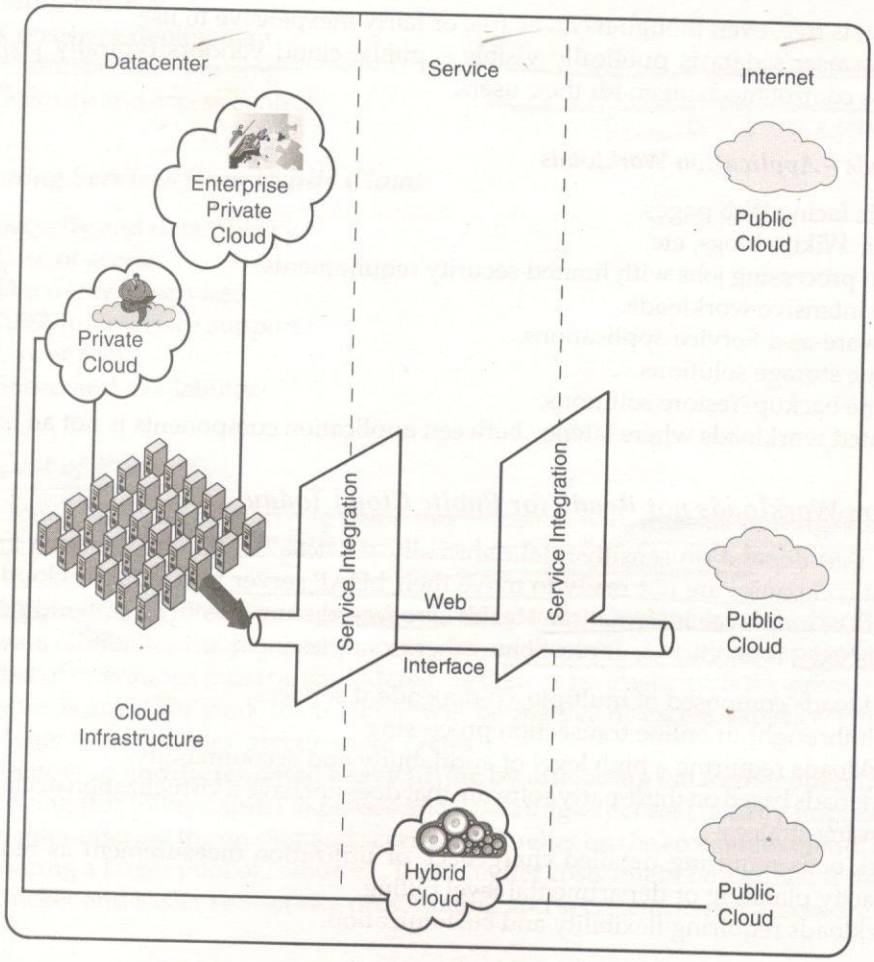


FIGURE 2.2 Private, public, and hybrid clouds.

moved to a public cloud – things such as search, e-commerce, and discreet business processes like sales force management.

There is not a one-size-fits-all model; in a number of cases, businesses may end up using all these models eventually, based on the business model for different services.

2.4.1 Public Clouds

Public cloud services are available to clients from a third-party service provider via the Internet. Public clouds provide an elastic, cost-effective means to deploy solutions and take care of deploying, managing, and securing the infrastructure. Companies can use it on demand, and with the pay-as-you-use option, it is much like utility consumption. Enterprises are able to offload commodity applications to third-party service providers (hosters).

The term 'public' does not mean

- That it is free, even though it can be free or fairly inexpensive to use.
- That a user's data is publically visible – public cloud vendors typically provide an access control mechanism for their users.

Public Clouds – Application Workloads

- Public facing Web pages.
- Public Wiki's, blogs, etc.
- Batch processing jobs with limited security requirements.
- Data intensive workloads.
- Software-as-a-Service applications.
- Online storage solutions.
- Online backup/restore solutions.
- Isolated workloads where latency between application components is not an issue.

Application Workloads not Ready for Public Cloud Today

Workloads that depend on sensitive data normally restricted to the organization are public today. Most companies are not ready to move their LDAP server into a public cloud because of sensitivity of employee information. Health care record – until security of cloud provider is well established – is another example. Some other examples include:

- Workloads composed of multiple, co-dependent services.
- High throughput online transaction processing.
- Workloads requiring a high level of auditability and accountability.
- Workloads based on third-party software that does not have a virtualization or cloud aware licensing strategy.
- Workloads requiring detailed chargeback or utilization measurement as required for capacity planning or departmental level billing.
- Workloads requiring flexibility and customization.

2.4.2 *Private Clouds*

Private clouds are deployments made inside the company's firewall (on-premise datacenters) and traditionally run by on-site servers. Private clouds offer some of the benefits of a public cloud computing environment, such as elastic on-demand capacity, self-service provisioning, and service-based access. They satisfy traditional requirements for greater control of the cloud infrastructure, improving security, and resiliency because user access and the networks used are restricted and designated.

Services in Private Cloud

This section highlights the services provided by private cloud and services consumed by public cloud specifically:

- Virtualization.
- Government and management.
- Multi-tenancy.
- Consistent deployment.
- Chargeback and pricing.
- Security and access control.

Consuming Services from Public Cloud

- Security and data privacy.
- Ease of access.
- Discovery of services.
- RESTful interface support.
- Lower cost.
- Speed and availability.

High 'Cost of Privacy'

Many experts believe that a private cloud implemented with internal hosting/running of the infrastructure makes it difficult to realize many key benefits of clouds, including:

- **Eliminating capital expenses and operating costs:** Ownership of the hardware or software eliminates the pay-per-use potential, as these must be upfront purchases. The full cost of operations must be shouldered, as there is no elasticity. If the private cloud hardware is sized for peak loads, there will be inefficient excess capacity. Otherwise, the owner faces complex procurement cycles.
- **Removing undifferentiated heavy lifting by offloading datacenter operations:** Utility pricing (for lower capital expenses and operating expenses) usually implies an outside vendor offering the on-demand services, and relies on the economies of multiple tenants sharing a larger pool of resources. These higher costs might be justified if the benefits of quicker and easier self-service provisioning and service-oriented access are large.

Private Clouds Provide more Control

In traditional security models, location implies ownership, which, in turn, implies control when security is location-specific. Then location, ownership, and control are aligned. Strong requirements for control and security usually drive a preference for a private cloud, where they own the cloud resources and control the location of those resources. For example, governments may not want their applications or data to reside outside certain borders. Clouds rely on virtualization, and in the public model, this loose coupling breaks the link between location and application, and this reduces the perceived ownership and control.

But control of information is not, in fact, dependent on total ownership or a fixed location. One example is public key encryption – the ownership of the key means control over the information without having to own the rest of the infrastructure. Control can be created over an untrusted infrastructure via a combination of encryption, contracts with service-level agreements, and by (contractually) imposing minimum security standards on the providers. Compliance is difficult outside traditional security models. As long as control through technology and contracts can be clearly demonstrated, it should be possible to make a public cloud computing environment as compliant and as secure as a privately owned facility. Auditors and regulators are continuously adapting to new technologies and business models.

There are two ends to the ownership spectrum – complete implementation ownership, and complete lack of ownership and control of implementation. There are many possible approaches in between, like partial control, shared ownership, etc. There are also different levels of limited access – specific departmental access, industry-only access, controlled partner access, etc.

2.4.3 Hybrid Clouds

A hybrid cloud is a combination of an interoperating public and private cloud. In this model, users typically outsource non-business-critical information and processing to the public cloud, while keeping business-critical services and data in their control. The hybrid model is used by both public and private clouds simultaneously, and is an intermediate step in the evolution process, providing businesses with an on-ramp from their current IT environment into the cloud.

It offers the best of both cloud worlds – the scale and convenience of a public cloud and the control and reliability of on-premises software and infrastructure – and lets them move fluidly between the two based on their needs. This model allows:

- Elasticity, which is the ability to scale capacity up or down in a matter of minutes, without owning the capital expense of the hardware or datacenter.
- Pay-as-you-go pricing.
- Network isolation and secure connectivity as if all the resources were in a privately owned datacenter.
- Gradually move to the public cloud configuration, replicate an entire datacenter, or move anywhere in between.

2.4.4 Community Clouds

A community cloud is controlled and used by a group of organizations that have shared interests, such as specific security requirements or a common mission. The members of the community share access to the data and applications in the cloud.

2.4.5 Shared Private Cloud

This is shared compute capacity with variable usage based pricing to business units that are based on service offerings, accounts datacenters and it requires an internal profit centre to take over or buy infrastructure made available through account consolidations.

2.4.6 Dedicated Private Cloud

Dedicated private cloud has IT Service Catalogue with dynamic provisioning. It depends on Standardized SO architectural assets that can be broadly deployed into new and existing accounts and is a lower cost model.

2.4.7 Dynamic Private Cloud

Dynamic private cloud allows client workloads to dynamically migrate to and from the compute cloud as needed. This model can be shared and dedicated. It delivers on the ultimate value of clouds. This is a very low management model with reliable SLAs and scalability.

2.4.8 Cloud Models Impact

Clouds will transform the IT industry. They will profoundly affect how we live and how businesses operate.

Cloud computing

- Provides massively scalable computing resources from anywhere.
- Simplifies service delivery.
- Provides rapid innovation.
- Provides dynamic platform for next generation datacenters.

Some say it is grids or utility computing or Software-as-a-Service, but it is all of those combined.

Public Clouds: Benefits

There are various ways to benefit from public clouds. Let us see some of the offering facilitated by public clouds:

- Lower barrier to entry/upfront investment.
- Offer self-service for rapid-start development.

- Deliver new pricing models for hardware, software, and service consumption.
- Increase or decrease capacity in minutes.
- Pursue new workloads and opportunities demo/sandbox, collaboration, prototypes.

Internal Private Clouds Drive Cost Savings

There are significant cost savings in implementing an internal private cloud versus a usual traditional infrastructure. With a traditional infrastructure, each server typically runs a single application and the hardware is sized to meet peak demands, which leads to very low average hardware utilization and high software costs due to the number of servers that are deployed and the lack of resource sharing. The internal private cloud uses virtualization on larger servers and leverages advanced service management capabilities to drive efficiency. Servers can be dynamically provisioned to adjust to workload changes and end-users can request the services they need through self-service portals, which drive automation.

Significant cost savings can be achieved by leveraging these capabilities to automate test and development environments. Automation drives down IT labour cost by automatically responding to changes in the environment and taking action before problems occur. Virtualization coupled with service management greatly improves server utilization and reduces software license costs since fewer machines need licenses. Automated provisioning and standardization allows systems to be provisioned in minutes by scripting the install process. In addition, end-users can now interface with IT through self-service portals to request services much like ATMs are leveraged to improve banking service. It can:

- Reduce IT labour cost by 50 percent in configuration, operations, management, and monitoring.
- Improve capital utilization by 75 percent, significantly reducing license costs.
- Lower administrative costs by 50 percent.
- Reduce end-user IT support costs by up to 40 percent.
- Reduce provisioning cycle times from weeks to minutes.
- Benefits of cloud economics with security within your firewall.
- Provide self-service for rapid-start development.
- Provide consistency of application environments.

2.4.9 *Savings and Cost Metrics*

Cloud computing's use of virtualization consolidates systems, which will drive reductions in hardware costs. This is often the initial appeal of funding virtualization projects.

Labour savings are even greater. Many companies still undertake the manual provisioning of IT systems, suffer long and costly delays while people wait for resources to become available, and distract highly skilled personnel from key project to focus on the mundane administration of systems. The automation of these tasks in a highly virtualized cloud environment can save significant labour costs while improving quality and productivity.

The total savings substantially off-set the small incremental increase in software costs that are usually necessary to deliver virtualization and the service management component that are elements of every cloud computing environment.

Cloud computing features two delivery models, private cloud computing and public cloud computing. Private cloud computing exists behind the firewall, while public cloud computing is accessed through the Internet. Cloud vendors believe that these three models – traditional IT, private cloud services, and public cloud services – will all co-exist as part of an overall strategy, based on application type and the business need that would dictate which model.

Hybrid clouds are services delivered to the end user that are composed of both private and public cloud computing elements.

2.4.10 *Commoditization in Cloud Computing*

When businesses started taking advantage of IT, the first organizations to computerize their business processes had significant gains over their competitors. As the IT field matured, the initial competitive benefits of computerization fell. Computerization then became a requirement just to stay on a level playing field. In essence, there is an increasing amount of IT that operates as a commodity.

For example, a paper products company needs a certain amount of unique IT to run its business and make it competitive. But it also runs a huge amount of commodity IT. The commodity technology takes time, money, people, and energy away from their business of producing quality paper products at a competitive price.

As executive management realizes it is operating a lot of commodity IT, which is not core to their competency, the debate shifts from whether cloud computing will take hold in the enterprise to a debate about how much of the organizational IT will be left internal, on-premise. IT functions should be evaluated, and a determination made as to which is a 'commodity' and which is not. Then determine where to place that function in the new IT organization.

2.5 SECURITY IN A PUBLIC CLOUD

Let us now discuss some of the security concerns that should be considered for the cloud deployments.

2.5.1 *Multi-Tenancy*

As long as the cloud provider builds its security to meet the higher-risk client, all of the lower-risk clients get better security than they would have normally. A bandage manufacturer may have a low risk of being a direct target of malfeasants, but a music label that is currently suing file sharers could have a high risk of being targeted by malfeasants. When both the bandage manufacturer and the music label use the same cloud (multi-tenancy), it is possible that attacks directed at the music label could affect the bandage manufacturer's infrastructure as well. So the cloud provider must design the security to meet the needs of the music label – and the bandage manufacture gets the benefits.

2.5.2 Security Assessment

Over time, organizations tend to relax their security posture. To combat a relaxation of security, the cloud provider should perform regular security assessments. The assessments should be done by someone who is experienced and able to identify issues and fix them.

The report should be provided to each client immediately after the assessment is performed so that the clients know the current state of the overall cloud's security.

2.5.3 Shared Risk

Sometimes, a cloud service provider may not be the cloud operator, but may be providing a value-added service on top of another cloud provider's service. For example, if a Software-as-a-Service (SaaS) provider needs infrastructure, it may make more sense to acquire that infrastructure from an Infrastructure-as-a-Service (IaaS) provider rather than building it. These cloud service provider tiers that are built by layering SaaS on top of IaaS, for example, can affect a cloud user's security. In this type of multi-tier service provider arrangement, each party shares the risk of security issues because the risk potentially affects all parties at all layers. This issue must be addressed by taking into consideration the architecture used by the cloud provider and working that information into the total risk mitigation plan.

Prepare strategy to prepare a list

2.5.4 Staff Security Screening

Most organizations employ contractors as part of their workforce. Cloud providers are no exception. As with regular employees, the contractors should go through a full background investigation comparable to the cloud user's own employees.

A cloud provider must be able to provide its policy on background checks and document that all of its employees have had a background check performed as per the policy. The contract between the user and cloud provider should bind the cloud provider to require the same level of due diligence with its contractors.

2.5.5 Distributed Datacenters

Disasters are a fact of life, and include hurricanes, tornadoes, landslides, earthquakes, and even fibre cuts.

In theory, a cloud-computing environment should be less prone to disasters because providers can provide an environment that is geographically distributed. But many organizations sign up for cloud computing services that are not geographically distributed, and therefore, they should require their provider to have a working and regularly tested disaster recovery plan, which includes SLAs.

Organizations that do contract for geographically diverse cloud services should test their cloud provider's ability to respond to a disaster on a regular basis.

2.5.6 Physical Security

Physical external threats should be analyzed carefully when choosing a cloud security provider. Do all of the cloud provider's facilities have the same levels of security? Are you being sold on the most secure facility with no guarantee that your data will actually reside there? Do the facilities have, at a minimum, a mantrap, card or biometric access, surveillance, an onsite guard, and a requirement that all guests be escorted and all non-guarded egress points be equipped with automatic alarms?

2.5.7 Policies

Any organization that says it has never had a security incident is either being deceptive or is unaware of the incidents it has had. It is, therefore, unrealistic to assume a cloud provider will never have an incident. Cloud providers should have incident response policies, and they should have procedures for every client that feed into their overall incident response plan.

2.5.8 Coding

All cloud providers still use in-house software, which may contain application bugs, so every organization should make sure that their cloud provider follows secure coding practices. Also, all codes should be written using a standard methodology that is documented and can be demonstrated to their customer.

2.5.9 Data Leakage

Data leakage has become one of the greatest organizational risks from a security standpoint. Virtually every government worldwide has regulations that mandate protections for certain data types.

The cloud provider should have the ability to map its policy to the security mandate users must comply with and discuss the issues. At a minimum, the data that falls under legislative mandates, or contractual obligation, should be encrypted while in flight and at rest. Further, an yearly risk assessment just on the data in question should be done to make sure the mitigations meet the need. The cloud provider also needs to have a policy that feeds into the security incident policy to deal with any data leakages that might happen.

2.6 PUBLIC VERSUS PRIVATE CLOUDS

A public cloud is a shared cloud computing infrastructure that anyone can access. It provides hardware and virtualization layers that are owned by the vendor and are shared between all customers. It is connected to the public Internet and presents an illusion of infinitely elastic resources.

Initially it does not require upfront capital investment in infrastructure. For consumption-based pricing, the user pays for resources used, allowing for capacity fluctuations over time.

Provisioning is applied through simple Web interface for self-service provisioning of infrastructure capacity. Potentially significant cost savings are possible from providers' economies of scale. Operating costs for the cloud are absorbed in the usage-based pricing. Separate provider has to be found (and paid for) to maintain the computing stack. Users have no say in SLAs or contractual terms and conditions. Sensitive data is shared beyond the corporate firewall. Distance may pose challenges with access performance and user application content for geographic locations. Support for operating system and application stacks may not address the needs of the business.

A private cloud is a cloud computing infrastructure owned by a single party. It provides hardware and virtualization layers that are owned by, or reserved for the business. It, therefore, presents an elastic but finite resource and may or may not be connected to the public Internet.

SLA's contractual terms & condition are negotiable

2.7 CLOUD INFRASTRUCTURE SELF-SERVICE

The cloud infrastructure has to be provisioned and paid for up-front in private clouds. Users pay for resources as used, allowing for capacity fluctuations over time. Self-service provisioning of infrastructure capacity is only possible up to a point in private clouds. Standard capacity planning and purchasing processes are required for major increases. For a large, enterprise-wide solution, some cost savings are possible from providers' economies of scale. The enterprise maintains ongoing operating costs for the cloud, and the cloud vendor may offer a fully managed service (for a price). SLAs and contractual terms and conditions are negotiable between the cloud vendors and customers to meet specific requirements. All data and secure information remains behind the corporate firewall and the option exists for close proximity to non-cloud datacenter resources or to offices if required for performance reasons for geographic locality. Private clouds can be designed for specific operating systems, applications, and use cases, unique to the business.

There is no clear 'right answer', and the choice of cloud model will depend on the application requirements. For example, a public cloud could be ideally suited for development and testing environments, where the ability to provision and decommission capacity at short notice is the primary consideration, while the requirements on SLAs are not particularly strict. Conversely, a private cloud could be more suitable for a production application where the capacity fluctuations are well understood, but security concerns are high.

Cloud computing employs a structured technique to holistically leverage IT industry best practices to uncover areas of relative strength and weakness across multiple IT domains (strategic alignment, computing system and storage, applications and data, processes, organization, finance/environment, and network) to determine readiness for a cloud computing deployment.

Infrastructure strategy and planning for cloud computing strategy gears the clients who are looking for assistance in understanding the business value that the cloud computing model can bring. It is designed to help the clients evaluate their readiness for cloud computing and possible cloud computing uses within their infrastructure. The goal is to develop a high-level vision strategy, value case, and roadmap for cloud computing.

Infrastructure strategy and planning for cloud computing employs a structured technique to holistically leverage IT industry best practices to uncover areas of relative strength and weakness across multiple IT domains to determine readiness for a cloud computing deployment. It is a business and IT executive initiatives to identify where and how cloud computing can drive business value.

2.7.1 *Infrastructure Strategy and Planning Features*

The strategy and planning has three major features:

- Assessment of the current environment to determine strengths, gaps, and readiness.
- Development of the value proposition for cloud computing in the enterprise.
- Strategy, planning, and roadmap to successfully implement the selected cloud delivery model.

Cloud-based systems have brought a new, scalable application delivery service model to the market. Cloud services promise to help reduce capital and operational costs while providing higher service levels. However, cloud services rely heavily on keeping the data and applications they are managing available at all times, and to restore operations quickly following any type of data disaster (database corruption, virus attack, hardware failure,local/regional disaster).

Cloud administrators need to ensure that a minimum of data is at risk by performing backups as frequently as possible, to meet stringent recovery point objectives. And downtime must be limited as well following an outage to meet strict recovery time objectives.

Emerging model where users can have access to applications or compute resources from anywhere with their connected devices through a simplified UI are best suitable alternatives for ease of use. Applications reside in massively scalable datacenters where compute resources can be dynamically provisioned and shared to achieve significant economies of scale. The 'pay-as-you-go' usage model enables users and companies to predict and manage expenses, reduce costs, and simplify operations better.

2.7.2 *The Path to Cloud Computing*

The path from simple virtualization to cloud computing occurs in five somewhat distinct stages.

Stage 1: Server Virtualization

Companies usually start virtualization as a consolidation attempt. The focal point tends to be on reducing capital expenses (like server, storage, and networks), reducing energy costs, and perhaps avoiding or delaying a datacenter build-out or move.

Stage 2: Distributed Virtualization

Once companies start down the virtualization way, and start to achieve capital expense improvements (like server, storage, and networks), the next focus tends to be on elasticity, operational improvements, rapidity, and organizing downtime more efficiently.

Stage 3: Private Cloud

Once processes are designed for alacrity and standards are in place to enable broad automation, the company is ready to look at introducing self-service capabilities based on the virtualization architecture.

Stage 4: Hybrid Cloud

Private clouds will not be the only answer for any enterprise. The self-service portals and interface introduced by private clouds should enable IT enterprises to leverage public cloud services when they make logic without affecting end users.

Stage 5: Public Cloud

Virtualization is not the must thing or are not the stepping stones before companies use public cloud services. Actually, some companies will attempt with cloud in the public cloud arena first, and use their lessons to establish private clouds for their enterprises.

2.8 SUMMARY

We have discussed several models for cloud computing, including private clouds (where the deployment is within the organization's firewall) and public clouds (where the application services and data are hosted by a third party outside the firewall). Consistent data availability and security is a critical success factor for any cloud deployment. Businesses need to ensure that data is adequately protected and can be restored in a timely fashion following any disruption.

CHAPTER 3

CLOUD AS A SERVICE

3.1 INTRODUCTION

In today's economy, many businesses are faced with the challenge of 'taking cost out' of their infrastructure while continuing to deliver new, innovative business services – basically they need to 'do more with less'. These days, many businesses face the necessities of a fast change to their IT infrastructure to manage requests during peak times. Organizations are dealing with IT resource optimization and lowering cost, and are looking for a way to manage these resources to meet such requirement. Also, they are trying to add rental-style capability to IT resource usage.

Cloud computing defines a new way to manage IT resources enabling self-service provisioning of IT resources, metering-style accounting based on use/time, automation of IT management in a standard process environment.

Cloud computing is a user experience and a business model. It is an emerging style of computing in which applications, data, and IT resources are provided as services to users over the network. Cloud computing is also an infrastructure management methodology. It is a way of managing large numbers of highly virtualized resources such that, from a management perspective, they resemble a single large resource, which can then be used to deliver services.

This chapter visualizes the different cloud models with respect to services. It also takes into account what service is all about and the different type of infrastructure services that can be offered as cloud as a service.

Common attributes of a cloud infrastructure are defined as:

- **Flexible pricing**, which means utility pricing, variable payments, pay-by-consumption; subscription models make pricing of IT services more flexible.
- **Elastic scaling**, which means resources scale up and down by large factors as the demand changes.
- **Rapid provisioning**, which means IT and network capacity and capabilities are – ideally automatically – rapidly provisioned using Internet standards without transferring ownership of resources.
- **Standardized offerings**, which mean uniform offerings readily available from a services catalogue on a metered basis.

There are two primary levers to achieve cost optimization – operating expense (Op-ex) and capital expense (Cap-ex) – and for many businesses, it is not just a question of lowering costs. It is also important to strike the right balance between Op-ex and Cap-Ex.

When you look at the different cloud types, the common terminology that comes up is 'as a service', with infrastructure as a service being the most basic type of service. It includes compute power, storage, and file systems as a service. At the next level, you have platform as a service, where a compute platform or middleware is provided.

At the next level, this is software as a service, and this is where you take software capability that would typically be in a package – like customer relationship management or

e-commerce – and you deliver that as a service. At the highest level is business process as a service. This is where a business can take a function that it considers to be a commodity, and not a differentiator, and just completely outsource it and buy it as a service. So cloud computing has really taken a hold because of the fact that you take virtualization, standardization, and automation, and drive this automated delivery of services at a reduced cost.

Cloud computing introduces the concept of 'IT-as-a-Service'. To support this service, the cloud infrastructure must deliver:

- **Abstraction:** Alleviate IT consumers from the operations of applications, allowing end-users to focus instead on the execution of high-value activities.
- **Virtualization:** Access to business services on-demand independent of location and resource constraints.
- **Dynamic allocation:** Dynamically provisions, configures, reconfigures, and de-provisions IT capability as and when needed, transparently and seamlessly.
- **Data management:** Fast, secure, reliable data access, and mobility, with integrated data protection and recovery management.

Because all data reside on the same shared storage systems, effective and efficient data and storage management become critical in a cloud deployment.

3.2 GAMUT OF CLOUD SOLUTIONS

Even within the cloud computing space, there is a spectrum of offering types. There are five commonly used categories.

- **Platform-as-a-Service (PaaS):** This is the provisioning of hardware and OS, frameworks and database, for which developers write custom applications. There will be restrictions on the type of software they can write, offset by built-in application scalability.
- **Software-as-a-Service (SaaS):** This is the provisioning of hardware, OS, and special-purpose software made available through the Internet.
- **Infrastructure-as-a-Service (IaaS):** This is the provisioning of hardware or virtual computers where the organization has control over the OS, thereby allowing the execution of arbitrary software.
- **Storage-as-a-Service (SaaS):** This is the provisioning of database-like services, billed on a utility computing basis, for example, per gigabyte per month.
- **Desktop-as-a-Service (DaaS):** This is the provisioning of the desktop environment either within a browser or as a terminal server.

The distinction between the five categories of cloud offering is not necessarily clear-cut. In particular, the transition from Infrastructure-as-a-Service to Platform-as-a-Service is a very gradual one.

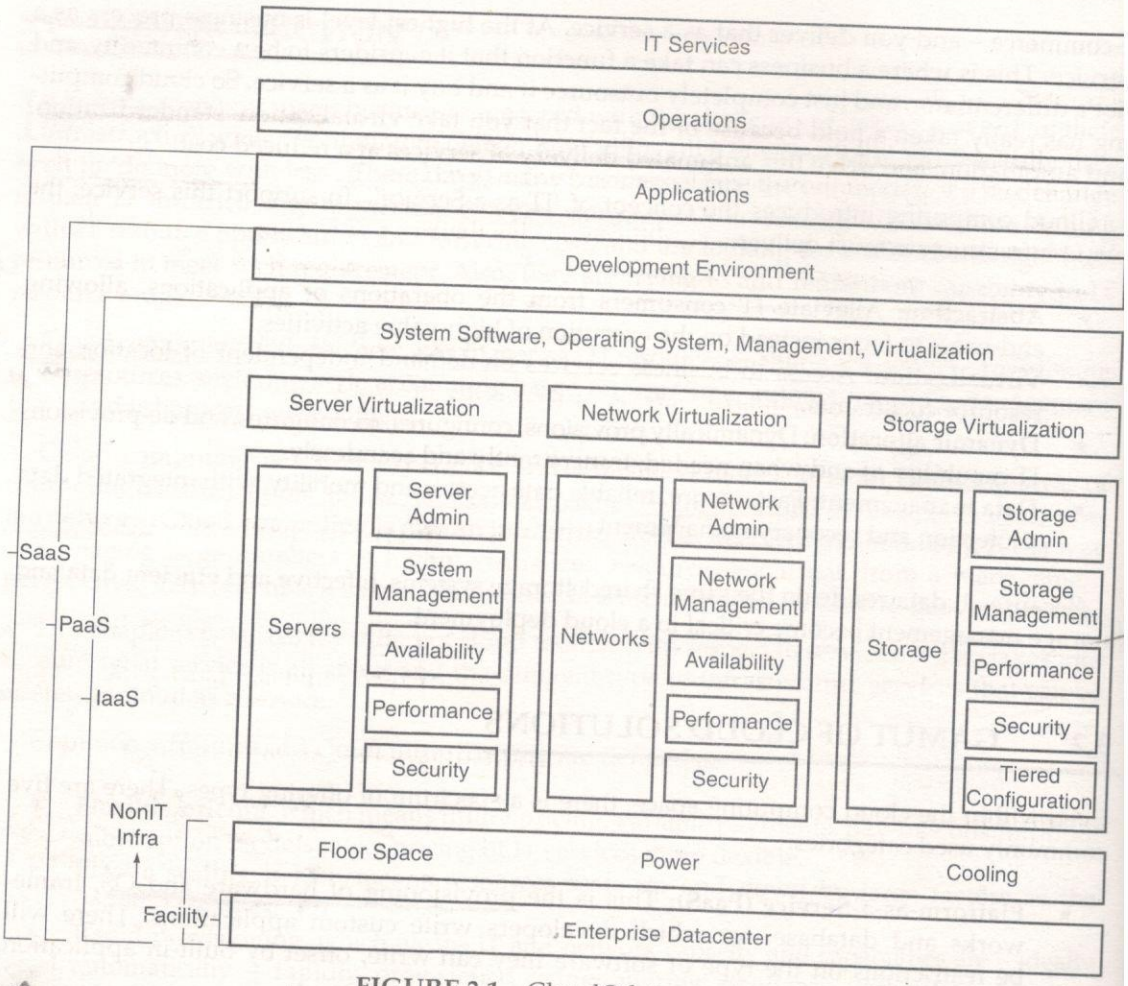


FIGURE 3.1 Cloud@datacenter.

Cost effective
Secure appliance
Charge back

3.2.1 *Platform-as-a-Service*

Instead of just offering applications over the Web in the form of Software-as-a-Service (SaaS), PaaS public cloud players are actually offering an entire Platform-as-a-Service (PaaS). They provide the foundation to build highly scalable and robust Web-based applications in the same way that the traditional operating systems like Windows and Linux have done in the past for software developers. What is very different about this model is that no longer is the platform itself 'sold' to the customer who is then responsible for running and maintaining it. In this model, it is this very operational capability of the platform hosting that is of primary value here (and that is how such platforms are typically billed). This has far-reaching implications to both the business models of PaaS vendors as well as their customers. One can use private clouds to speed application deployments for fast deployment in minutes. It will help tracking usage for the chargeback and will give the option of cost effective and secure appliance.

In order to optimize deployments, many organizations are looking to extend SOA to cloud services.

Cloud capabilities can improve the productivity of your development and test teams to roll out new applications and SOA services faster and reduce application backlog. It provides a catalogue of virtual images, and patterns all ready for immediate use. Patterns define a cluster of servers working together.

PaaS saves costs by reducing upfront software licensing and infrastructure costs, and by reducing ongoing operational costs for development, testing, and hosting environments.

PaaS significantly improves development productivity by removing the challenges of integration with services such as database, middleware, web frameworks, security, and virtualization. Software development and delivery times are shortened since software development and testing are performed on a single PaaS platform. There is no need to maintain separate development and test environments.

PaaS fosters collaboration among developers and also simplifies software project management. This is especially beneficial to enterprises that have outsourced their software development.

There is a challenge for tight binding of the applications with the platform which makes portability across vendors extremely difficult. PaaS in general is still maturing, and the full benefits of componentization and collaboration between services is still to be demonstrated. PaaS offerings lack the functionality needed for converting legacy applications into full fledged cloud services.

SaaS, PaaS, and IaaS suit different target audiences. SaaS is intended to simplify the provision of specific business services. PaaS provides a software development environment that enables rapid deployment of new applications. IaaS provides a managed environment into which existing applications and services can be migrated to reduce operational costs.

3.2.2 *Software-as-a-Service*

SaaS saves costs by removing the effort of development, maintenance, and delivery of software; eliminating up-front software licensing and infrastructure costs; and reducing ongoing operational costs for support, maintenance, and administration.

The time to build and deploy a new service is much shorter than for traditional software development. By transferring the management and software support to a vendor, internal IT staff can focus more on higher-value activities.

Applications that require extensive customization are not good candidates for SaaS. Typically, this includes most complex core business applications that will not be the best suit for SaaS.

There are also issues involved in moving to SaaS. Moving applications to the Internet cloud might require upgrades to the local network infrastructure to handle an increase in network bandwidth usage. Normally only one version of the software platform will be provided. Therefore, businesses are obliged to upgrade to the latest software versions on the vendor's schedule. This could introduce compatibility problems between different vendor offerings.

3.2.3 *Infrastructure-as-a-Service*

IaaS saves costs by eliminating the need to over-provision computing resources to be able to handle peaks in demand. Resources dynamically scale up and down as required, reducing capital expenditure on infrastructure and ongoing operational costs for support, maintenance, and administration. Organizations can massively increase their datacenter resources without significantly increasing the number of people needed to support it (Figure 3.2).

The time required to provision new infrastructure resources is reduced from typically months to just minutes – the time required to add the requirements to an online shopping cart, submit it, and have it approved. IaaS platforms are generally open platforms, supporting a wide range of operating systems and frameworks. This minimizes the risk of vendor lock-in.

Infrastructure resources are leased on a pay-as-you-go basis, according to the hours of usage. Applications that need to run 24×7 may not be cost-effective. To benefit from the dynamic scaling capabilities, applications have to be designed to scale and execute on the vendor's infrastructure. There can be integration challenges with third-party software packages. This should improve over time, however, as and when independent software vendors (ISVs) adopt cloud licensing models and offer standardized APIs to their products.

SaaS is considered to be considerably more mature as a cloud offering than PaaS or IaaS. Even then, it is mainly Small & Medium Businesses that have adopted cloud services. Adoption by the larger enterprises is still extremely low. A perception of cloud services as a high-risk technology option has led the large organizations to restrict the use of cloud services

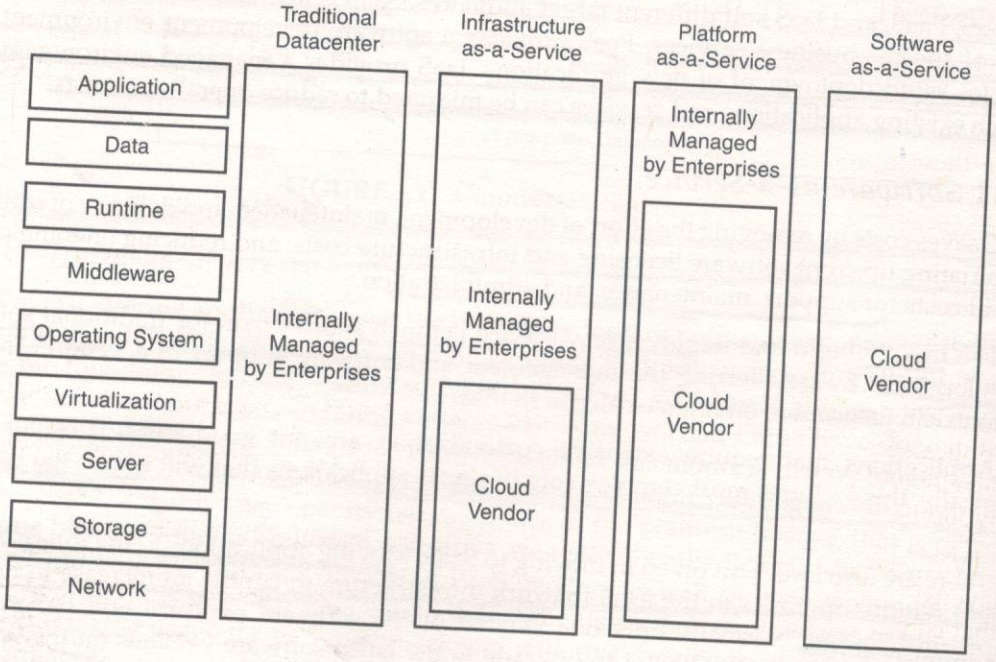
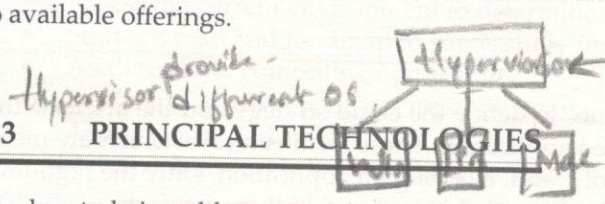


FIGURE 3.2 Cloud taxonomy.

to a limited number of projects. PaaS is a more sophisticated service platform, and is still an emerging product. It will need to stabilize and mature before developers can use it for extensive building of new SaaS applications. For IaaS, the entry of large vendors such as Amazon and other cloud vendors is driving up the maturity of the offering rather quickly, with Amazon leading efforts to define the IaaS market.

There are now hundreds of vendors offering some flavour of cloud computing. Other vendors have attached themselves to the 'Cloud bandwagon' by providing ancillary services to available offerings.



3.3 PRINCIPAL TECHNOLOGIES

The key to being able to provide the dynamic cloud infrastructure is the virtualization layer that sits between the cloud instances and the physical hardware it runs on. The platform virtualization software – the hypervisor – allows multiple operating system instances to run as guests on the same server.

The main drivers for cloud computing are *cost*, *agility*, and *time to market*. By building cloud infrastructures using any cloud orchestrator and provisioning engine one can realize cost savings and improve time to market. This sits on top of the virtualization layer working on network, server, and storage. It is a layer of software that (a) interacts with multiple servers, (b) enables IT departments to pool resources together across servers, and (c) defines standardized tiers of services called *virtual compute centres*. This helps break down infrastructure silos and drives sharing of infrastructure.

Using cloud orchestrator and provisioning engine, IT departments can define organizations and on-board users that can share the underlying cloud infrastructure in a secure multi-tenant fashion. IT can then create standardized collections of VMs and set policies on how users can use these VMs. Users can login into orchestrator and self-provision respective workloads which IT has setup already. This effectively removes IT involvement each time users require infrastructure-enabling agility and faster time to market for applications.

These orchestrators also transport with the API which allow cloud administrator and users to interrelate with the cloud infrastructure in a systematic way. In addition, cloud orchestrator and provisioning engine allows writing workflows to automate creation of cloud infrastructure.

The increased pooling and sharing of resources, self-provisioning, and increased automation deliver greater cost savings in IT infrastructure, agility, and faster time to market for applications. One can deliver cloud benefits for today's applications and for applications developed in the future as cloud orchestrator and provisioning engine builds the top of the hypervisor layer.

Virtualization is the foundation for cloud. It consists of physical hardware with hypervisor layered on top of it. Cloud orchestrator and provisioning engine consists of one or more cells that communicate with a single database and offer a web portal. Using the web portal, cloud administrators create cloud infrastructure resources and users' self-provision cloud infrastructure resources in a secure multi-tenant fashion, thus enabling Infrastructure-as-a-Service (IaaS).

Chargeback and metering is a key piece of the on-premise cloud solution. This server ta
to own database, server database, and cloud orchestrator and provisioning engine databa
and allows to associate costs with the cloud and generate usage and billing reports. This shou
also integrate with workflow systems, LDAPs, approval process, etc. to provide the lifecy
management of the cloud environment.

3.4 CLOUD STRATEGY

This section provides a high-level guidance to define the cloud strategy and the artefacts tha
capture the architecture of a cloud-enabled application. These architectural artefacts are mean
for the implementation planning phase of cloud, enabling an application. Only the high-leve
architecture of the system that is to be cloud-enabled is captured in these artefacts.

The implementation planning phase of cloud enables an application lying between the
business strategy definition for the adoption of cloud and the design, development, and imple-
mentation phases of the application that is being implemented to be offered on the cloud plat-
form. It takes care of linking the business strategy that is defined for a business to adopt a
cloud-based strategy and the IT requirements for the applications on the cloud that are needed
to support this strategy. So, this critical piece of the implementation planning translates the
business intent to a set of IT requirements for the cloud-based application, deriving the high-
level structures of the cloud-based application and defining a roadmap for the implementation
of the application.

So, the primary input for the cloud implementation planning phase comes from the cloud
strategy for the business that is driving the cloud-based implementation of one or more
applications on the cloud.

The key steps in cloud implementation planning are as follows:

- Understand cloud strategy.
- Define cloud application requirements.
- Assess cloud readiness.
- Define high-level cloud architecture.
- Identifying change management requirements.
- Develop roadmap and implementation plan.

Of these phases, the artefacts defined here relate to the phase of 'Defining the high level
cloud architecture'. In this phase, the high-level structure of the cloud-based application is
defined from the inputs of the previous step, the 'Define cloud application requirements'.

The first step in the architecture development is the usage of existing asset analysis to under-
stand which of the components, including components of the existing application, can be used
for building the application and how they can be leveraged in the cloud environment. The next
step is to derive the high-level structure of the solution from the IT requirements for cloud in
the form of the following artefacts. The non-functional characteristics that the application must
support and deliver are captured in the artefact 'Non-Functional Requirements'.

Infrastructure strategy and planning for cloud computing helps you develop a cloud strategy, plan, and roadmap:

- Business and IT executive workshop to identify where and how cloud computing can drive business value.
- Develop the value proposition for cloud computing in the enterprise.
- Identify priority of workloads to migrate to cloud.
- Assess the current environment to determine strengths, gaps, and readiness.
- Strategy, plan and roadmap to successfully implement the selected cloud.
- Analyse cloud computing opportunity.
- Analyse IT environment and capability gap.
- Assess cloud readiness.
- Develop high-level cloud roadmap and value proposition.

This helps to deploy the cloud deployment with following benefits:

- **Reduced risk and faster deployment:** It leverages cloud vendor assets, skills, and experience to reduce risk. It accelerates development and implementation by identifying the gaps, activities, and risks and defines mitigation strategies within an implementation roadmap.
- **Improve service:** It identifies the optimal delivery model mix and prioritizes the workloads to migrate to cloud to achieve your business and IT objectives.
- **Lower cost:** It identifies opportunities to reduce capital and operating expense across the infrastructure.

3.5 CLOUD DESIGN AND IMPLEMENTATION USING SOA

Service-Oriented Architecture (SOA) is a very useful architectural style for implementing applications in the cloud. Adoption of SOA would provide the best way to leverage and consume the application services provided by the cloud. A cloud-based application consists of many granular coarse-grained services offered on the cloud. These cloud offerings may in turn integrate and leverage services and systems from different environments. The coarse-grained services that the cloud offerings leverage can come from traditional IT environment in the form of standard services already provided and are internal to the cloud.

Standardization across these different environments is not possible, and hence, they mostly consist of heterogeneous environments. So a cloud service offering should be based on open standards that can be consumed and leveraged by this environment. Language- and platform-independent services can be provided using standards-based, platform-agnostic SOA architecture, with support for appropriate industry and technology standards.

The cloud-based application services consist of a coherent set of business processes that are aligned to the business boundaries, provide a coherent, integrated set of operations that adheres to the business intent and provides value to the users. They also comprise the consumable interfaces whether they are user interfaces on different devices, coarse-grained Web service interfaces, feeds, or widgets.

Map the cloud application services to the business processes that they consist of. These business processes could be at different levels. Drill down the processes to the levels that make business sense.

These processes, along with business strategy, the modular business alignment model, and the process scenarios to be supported form the input to Service-Oriented Modeling and Architecture (SOMA) methodology. SOMA is applied with a meet-in-the-middle approach. The business processes and business strategy along with business goals stated are used to arrive at the service portfolio using SOMA. While deriving the services, a bottoms-up approach is also taken, taking into consideration the existing assets, which consist of the existing application components and services available on the public clouds and internal to the cloud as well as industry cloud component maps.

3.5.1 *Architecture Overview*

The purpose of the architecture overview artefact in a cloud-based implementation is to communicate to the sponsor and external stakeholders a conceptual understanding of the architectural goals of the cloud implementation. It offers a layered conceptual model of the application services to be cloud-enabled and provides a high-level vision of the cloud architecture and its scope to developers. It is easy to explore and evaluate alternative architectural options for the cloud implementation. This stage enables early recognition and validation of the implications of the cloud-based architectural approach and facilitates effective communication between different communities of stakeholders and developers.

This artefact provides an overview of the main conceptual elements of the cloud implementation and relationships within and outside the cloud infrastructure, which may include other cloud and non-cloud environments, candidate offerings, cloud components, nodes, connections, data stores, users, external systems, and technical components to support requirements. As such, it represents the governing ideas and candidate building blocks of the cloud implementation.

This artefact can also provide key stakeholders with the first high-level view of the cloud architecture landscape of their transformed cloud-based implementation.

Typically, the artefact is produced over multiple iterations and as the project moves through the solution definition and design phases, the conceptual models get clearer and this document is kept up-to-date and governed to always have the conceptual elements current.

An architecture summary depiction represents the governing ideas and candidate building blocks of a cloud-based offering and enterprise architecture. It also provides an overview of the main conceptual elements and relationships in the architecture. The main purpose of such a depiction is to communicate a simple, brief, clear, and understandable overview of the target IT system.

At the enterprise-level, an architecture summary depiction is often produced as part of an overall IT strategy to move to a cloud-based offering model. In this instance, it is used to describe the vision of the business and IT capabilities required by an organization that needs the offering to be hosted on the cloud. It provides an overview of the main offerings and the relationships with other offerings, external systems, components, nodes, connections, data stores, users, external systems, and a definition of the key characteristics and requirements.

At a system- and component-level, the architecture summary depiction is developed very early in the project (possibly pre-proposal), and influences the initial cloud hosting vision with the component model and operational model. It is intended that design commitments be based on this conceptual overview as the (more detailed) component model and operational model are developed and validated. Subsequently, the component model and the operational model are the primary models used for implementation activities, while their compliance with the architecture summary depiction is maintained continually. Changes to the architecture summary depiction are made using a governance process.

For most infrastructure cloud engagements, the project scope is something less than the client's enterprise architecture. Therefore, the depiction will represent the governing ideas and candidate building blocks of the offerings that are to be hosted on the cloud and represents the focus of the engagement. There can be, then, multiple views of the cloud-based IT environment represented by architecture summary depiction: the current and future view, views by waves of transformation or views by types of clouds, for each solution alternative.

3.6 CONCEPTUAL CLOUD MODEL

So far, we have discussed the different service paradigms for cloud deployment. Now let us discuss the conceptual cloud model based on cloud services.

The conceptual cloud model describes the structure of the cloud-based services as a system in terms of its software components with their responsibilities, interfaces, relationships, and the way they collaborate to deliver the required functionality. The cloud component model for implementation planning is specified at the conceptual level.

The highest level of the conceptual model is the set of cloud-based offerings that make up the cloud-based business solution. These highest-level conceptual components are derived based on the business intent and business functionality. These conceptual components, which form the offerings, align tightly to the business intent that initiated the creation of the cloud-based offerings.

The conceptual offerings that provide the cloud-based implementation of the solution can be further broken down and depicted in a layered composition. The next level of conceptual component model broken down below the offering consists of the following elements:

- High-level service components that form the services provided by the offerings.
- The resources that support the cloud services.
- The technical components that provide the technical underpinnings of the cloud service and support the non-functional needs of the cloud application.
- External and internal services that are leveraged by the cloud application services.

The high-level conceptual component model is the main artefact that provides an abstract view of the design of a cloud application to business stakeholders. This abstract view describes how the business needs will be met by the cloud application components without delving into their technical details. Components identified can be decomposed into further layered conceptual component structures to convey further details of the respective components (Figure 3.3).

FIGURE 3.3 Cloud model.

The cloud conceptual component model should contain the following elements:

- The conceptual structure of the cloud application.
- The dynamic interactions and dependencies between various conceptual components.
- The components that comprise the cloud services provided by the application, each of which may be made up of sub-components.

At this level, the conceptual components are not converted to physical components but retained at the conceptual level only.

3.6.1 Cloud Application Security and Privacy Principles

The cloud application security principles contain the high-level guidance to cover the security and privacy characteristics identified in the Information Asset Profile. This is a summary document of the information asset required for the architecture. Cloud application security principles deliverable contains a number of security principles that occur in a typical application. Typically, quite a number of the characteristics identified in the information asset profile effort will relate to existing enterprise security elements in the client.

In addition to the elements from the existing enterprise security program, consider the following areas that tend to change as part of moving to the cloud environment.

3.6.2 Governance

How will decisions like different SLAs be made between cloud vendors and cloud customers for the cloud environment given the new stakeholders? These decisions are typically made by either another infrastructure group within the client IT organization or an external vendor who provides cloud infrastructure management for either a managed private cloud or public cloud environment.

Authentication and Access Control

As an application moves to the cloud, the methods used to authenticate and authorize users require examination. Existing methods may require additional support to work in the cloud environment, such as providing connectivity to an existing server, and may require significant extension to accommodate the requirements of an external vendor providing infrastructure support.

Data Protection

With the move to a cloud environment, the existing methods of data protection – protecting the data from disclosure or modification both on the network and on a storage medium – are likely to require change with the move to virtualized storage and with the introduction of additional infrastructure administrators, often from a vendor organization.

Logging and Alerting

Logging is the ability to tie actions to an individual. Alerting is the method of recognizing activities that may indicate a malicious act and bringing those to the attention of the security

staff. With the move to a cloud infrastructure, with different network and additional administrators, often from an external vendor, tying actions to specific individuals requires additional attention.

The focus in this work product is to provide the broad guidance as to the security and privacy elements that must be present in and around the cloud application.

3.7 CLOUD SERVICE DEFINED

This section highlights the aspects of service, its scope, and cloud-based different platform integration and deployment services.

3.7.1 *Service Definitions*

Let us begin with a cursory glance on what 'Service' is.

- **Service:** A specific IT deliverable that provides customer value. It is measurable in customer terms and provides the basis for doing business with the customer. It is delivered through a series of processes and/or activities.
- **Services Portfolio:** The collection of services provided by IT that in their aggregation represents all the 'value add' activities performed by IT.
- **Service Component:** A logically grouped set of activities that represent part of a service that touches the customer. Service components are grouped together to create deliverables that form the basis for doing business with customers.
- **Service Owner:** The individual accountable for ensuring the customer receives the identified value of the service. They take the customer's end-to-end view of IT activities by working with process owners to ensure all the required delivery components fit together smoothly.
- **Process:** A collection of related activities that take inputs, transforms them, and produces outputs that support an enterprise goal.
- **Enablers:** The decomposed components of a service (process, organization, technology) that are combined to create *Service Deliverables*. The collection of activities (and their supporting roles and technology) become the workflow needed for *Service Delivery*.
- **Service Level Agreements:** A grouping of Services or Service Components that have had specific delivery commitments and roles identified with the customer. SLAs can be grouped together in different ways to represent various products. Examples would include things like e-mail and service delivery.
- **Service Level Management:** Service Level Management governs the planning, coordinating, drafting, agreeing, monitoring, and reporting on Service Level Agreements (SLAs), and the on-going review of service achievements to ensure that the required and cost-justifiable service quality is agreed to, maintained, or where necessary improved. SLAs provide the basis for managing the relationship between the provider and the customer. Service Level Management is essential in any organization so that the level of IT Service needed to support the business can be determined and monitoring can be

initiated to identify whether the required service levels are being achieved – and if not, why not.

- **Service Level Management Objectives:** SLA objective is to maintain and improve IT Service quality, through a constant cycle of agreeing, monitoring, and reporting upon IT service achievements and instigation of actions to eradicate poor service – in line with business or cost justification. It develops a better relationship between IT and its customers.

SLAs should be established for all IT services being provided. Underpinning Contracts (UCs) and Operational Level Agreements (OLAs) should also be in place with those suppliers (external and internal) upon whom the delivery of service is dependent.

SLAs are likely to be a service... if the 'what' is separated from the 'how', and we can change the underlying assets – processes, technologies, data – as well as suppliers, but still provide the promised business outcomes and value. It is a service if its description and design highlight separate roles for customers, users, providers, and suppliers, and it is offered with a pre-defined value proposition stating specific price points and performance metrics tied to business (not just IT operational) outcomes. It is something that makes sense as a customer/user-selectable item, with corresponding service requests available in a menu or service catalogue.

The facts here are startling – inefficiency is prolific – clearly, progress is needed.

3.7.2 *Services Scope Overview*

Platform Integration and Deployment Services

These provide a set of project services for the planning, design, procurement, assembly and integration, site installation, and project management of the deployment mainstream and special-purpose end-user devices, and also includes a number of asset lifecycle services.

It integrates and customizes multiple devices – including PCs, wireless and mobile devices, kiosks, ATMs, point-of-sale (POS) devices, and printers – to end-user specifications. These services utilize a factory approach for the off-customer-site build and integration services, using build and integration centres around the globe.

Software Platform Management Services

It provides a set of project and annuity services to manage end-user software platforms, including image development and management, application software packaging and distribution, and services to manage the availability of the end-user platform proactively. This includes services to design and migrate end-user platforms, for example, Microsoft XP to Vista, or Linux. It utilizes a factory approach for the off-customer-site platform management services, using management and configuration centres.

3.7.3 *Platform Integration and Deployment Component Services*

This section discusses some of the very important platform integration and deployment component services.

Order Management

This service handles the procurement of hardware and/or software on behalf of the customer (regardless of asset ownership), and the fulfilment (delivery) of that hardware and software to the central build centre prior to Platform Build and Pre-Load.

Warehousing and Stock Management

This gives provision of central warehousing facilities to store hardware and other agreed components before and after Build and PDP and before shipment to site. It provides services for receiving and warehousing, and ensures that inventory is sufficiently maintained and protected while in storage.

Platform Build and Test

This service provides services to build, integrate, customize, prepare, and test the hardware and software platform before shipment to the customer site or end-user location.

Base Backup

This service provides a base backup to be taken during platform pre-build (assumes the software platform supports this feature).

Data and Personality Migration

It migrates data and personality settings (e.g., desktop wallpaper, Internet Favourites, Desktop layout) from original to replacement platform.

Asset Tagging and Custom Labelling

This service provides for custom asset tagging of hardware components during Build and pre-delivery preparation at the central build facility.

Asset Inventory Update and Report

This service includes adding any new asset to the customers or managed asset database.

Logistics and Delivery

This service includes the packaging and shipment to site of the user hardware platform following Platform Build and Test services.

Installation

This service provides the deployment of the platform into the customer's operational environment. This could be either at the end-users' desk, or agreed deployment centre location for machines that are to be collected or deployed by the customer.

Extended Project Management

This is an extension to the base project management services that cloud uses to manage internal functions and activities. This service extends project management scope to cover overall management of the platform deployment program, including customer and customer third-party resources and activities.

Platform Removal and Return

This includes decommission of platform from customer location, and return to cloud vendors for refurbishment or disposal (but not including for refurbishment or disposal activities).

Asset Refurbishment

This service checks for suitability of the hardware platform, refurbishment, and upgrade as appropriate, and integration into the deployment process.

Asset and Data Disposal

This service provides removal of sensitive data from the hardware platform to varying levels of security (e.g., to department of defence standards), as well as safe environmental disposal to international standards and/or asset value recovery.

Emergency Replacement

This service provides for the emergency replacement of like for like hardware platforms (pre-built to the customers standards) to agreed service levels.

Software Platform Management Services

This section explains the software platform management services in detail.

Software Platform Design Consulting

These services help the client understand the business and technology needs for a new software platform, and provide the design specification.

Software Platform Creation and Customisation

This service is to create and test a software platform to support the needs of the business.

Software Platform Support and Maintenance

It is for ongoing support and maintenance of the client's software platform, including platform updates and management, and ongoing problem support.

Application Scripting

This is the creation of application software unattended installation scripts to accommodate the customer environment. This service can include new software applications and updates to existing packages.

Application Discovery

This service is to discover what applications are in use across the organization.

Application Portfolio Management

This service is to help the clients manage their application portfolio.

Software Delivery

It services the schedule and transport application packages to target end points. This service includes the capability to PUSH software out to user PCs from a central distribution and/or the ability for the users to PULL Software to their PCs using a Web interface.

Antivirus Management

These are the services to manage the delivery of antivirus signature files. It provides real-time protection from malicious virus attacks that can cause potentially disastrous system.

Patch Management

It services to ensure end-user devices have the latest patch releases installed. It provides real-time deployment of critical Operating System patches to help protect against flaws and vulnerabilities.

Health Check Services

These services ensure that end-user devices are in a good state of health of the system, like infrastructure and PC checks. It performs remote monitoring of supported workstations for critical hardware alerts and initiates scripted responses as alerts are received. Examples are identification of low memory, detection of spyware, identify low disk space.

Compliance Services

These services are to ensure that end-user devices are compliant with client standards. They perform remote monitoring of supported workstations to perform activities such as detecting peer-to-peer software detection of games, etc.

Until there is an agreement to industry standard definition of what an IT service is, you must agree with the client on a definition of 'service'. So, we will consider different type of the services visibility with respect to customer.

External Services

These services are visible and seen by the customer, and include business services (business intelligence, logistics, receiving orders, marketing services, invoicing, accounting, etc.) and user services (desktop support, maintenance, education, etc.).

Internal Services

These services are invisible or less visible to the customer, but essential to the delivery of IT services. They include infrastructure services (hosting services, storage, availability, data

retention, or recovery) and network services (network, remote access, mobile or wireless services, etc.). It takes care of application services (integration, testing, design, maintenance, optimization, etc.).

User-Initiated Service Request

It includes service request handling in incident management, for example, the service request progresses through its lifecycle exactly as an incident. Most companies separate user service requests and incidents. For example, the service requests follow a different process and use a different tool to track it throughout its lifecycle. Significant IT workload is responding to user-initiated requests for some work to be done.

Users request services for which their businesses have already contracted with a service provider, and to which they are already entitled. Some people have referred to the list of services from which a user can order services as a service catalogue. Perhaps it is a 'User Services Catalogue'. These service requests and this type of listing of services are appropriately offered through the single point of contact for users of IT services, the service desk.

Customer-Initiated Service Request – A Service Catalogue-Based Request

It includes the concept of a service catalogue as a list of services that the customer can order. The customer is the one who pays for services. When customers order a service, their users are entitled to receive services under that agreement. Each individual request made by a user is a user initiated service request. The user is entitled to services that their business has ordered through a service catalogue request. This type of request is from the customer to some account rep or 'business relationship manager' who responds to this request and initiates the service provisioning for that customer.

3.8 SUMMARY

This chapter visualizes the different cloud models with respect to services. It also takes into account what service is all about and the different types of infrastructure services that can be offered as cloud as a service.

CHAPTER 4

CLOUD SOLUTIONS

4.1 INTRODUCTION

Cloud environment presents an opportunity to enhance the user experience by providi a broader communication path for reaching out to the user or for providing a series business services to the user via the application features.

Deploying the application to the cloud is somewhat different since the deployment proce will not be done locally within the enterprise and the existence of the provisioned image and series of deployment steps is needed to deploy the application and validate the deployment

Development and testing environments are readily available within the cloud environmer The advantages of these environments, especially from a costing perspective, are numerous there is no need to purchase and deploy servers within the normal enterprise environment. If POC was being developed and project was cancelled, no software, hardware, or even deve opment tools would have to purchased, only to be thrown away later as the cloud suppor development and testing of applications.

4.1.1 Cloud Application Planning

The design and development of cloud applications requires many unique considerations:

- Business functions.
- Application architecture.
- Security for cloud computing.
- Cloud delivery model.
- User experience.
- Development, testing, and run-time environments.

Application architecture is selected through some sort of criteria evaluation.

The key thing to talk about from a security aspect is the enhancements to the existing security model where data protection and isolation of the data from other areas of the cloud environment. Encryption is one possibility to further enhance the security model whereas the enterprise would not necessary invoke that option. Other aspect of security would be to further authenticate and authorize users of the application and the services the users have been entitled to use.

4.1.2 Cloud Business and Operational Support Services (BSS and OSS)

Business Support Services (BSS) are the components that cloud operators use to run their business operations. The term BSS applies to service providers in all sectors such as utility providers. Typical types of activities that count as part of BSS are taking customer orders, managing customer data, managing order data, billing, rating, and offering services.

Operational Support Services (OSS) are computer systems used by cloud service providers. The term OSS most frequently describes 'network systems' dealing with the network itself, supporting processes such as maintaining network inventory, provisioning services, configuring network components, and managing faults.

BSS and OSS components need to be externalized so that the supporting services of the application being transformed to the cloud environment can capitalize on the various functions the OSS and BSS provides. For example, provisioning can be adapted to support the applications provisioning requirements instead of creating self-provisioning from scratch. The ability to tap into the monitoring, metering events and keep track of the activity within the cloud environment can assist the application to continue to provide the service levels and quality of service. Each of the OSS and BSS offered by the cloud environment will continue to support the application and the consumers of the application in maintaining the key characteristics of cloud computing.

The cloud application architecture brings together the business services, security, infrastructure, and integration required for an optimal solution. Cloud services represent any type of IT capability that is provided by the cloud service provider to cloud service consumers. Typical categories of cloud services are infrastructure, platform, software, or business process services. In contrast to traditional IT services, cloud services have attributes associated with cloud computing, such as a pay-per-use model, self-service usage, flexible scaling, and shared of underlying IT resources.

The cloud vendor is responsible for delivering instances of cloud services of any category to cloud service consumers, the ongoing management of all cloud service instances from a provider perspective, and allowing cloud service consumers to manage their cloud service instances in a self-service fashion. The technical aspects of a cloud service are captured in a service template, which is also the artefact that describes how the OSS capabilities of the cloud vendor are exploited within the context of the respective cloud service.

For most cloud services, specific software are required for implementing cloud service specifics: For IaaS, these are typically hypervisors installed on the managed infrastructure; for PaaS, this would a multi-tenancy enabled middleware platform; for SaaS, a multi-tenancy enabled end-user application; and for BPaaS, multi-tenancy enabled business process engine. Depending on the nature of the respective cloud service, the notion of a cloud service instance represents different entities.

Cloud services can be built on top of each other, for example, a software service could consume a platform or infrastructure service as its basis, and a platform service could consume an infrastructure service as its foundation. However, this is not required, that is, a software service could also directly be built on top of 'traditional' infrastructure, clearly inheriting all constraints associated with such an infrastructure. In general, basic cloud architectural postulates to share as much as possible across cloud services with respect to management platform and underlying infrastructure. However, it does not require to only having one single, fully homogeneous infrastructure – of course, this would be the ideal goal, but given different infrastructure requirements, this is not possible. For example, if a particular cloud service has very specific infrastructure needs, it is clearly allowed to run this cloud service on a dedicated infrastructure (e.g., the Google search engine or HPC cloud services would always run on a purpose-built physical infrastructure for performance and efficiency reasons; they wouldn't run virtualized compute cloud service).

In the context of building cloud services on top of each other, it is important to distinguish the sharing of a common OSS/BSS structure across multiple cloud services and the usage of the actual cloud service capability by another one.

In any case, each cloud service offered by a cloud service provider is 'known' to the BSS and OSS of the cloud vendors. Consequently, a cloud service provider offers cloud services as a result of very conscious business decisions, since taking a cloud service to market must be supported by a corresponding solid business model and investments for the development and operations of the cloud service.

4.2 CLOUD ECOSYSTEM

It is very important to understand the relationship between a cloud service and the artefacts that can be developed based on and within the boundaries of an ecosystem-focused IaaS or PaaS cloud service. Bringing any cloud service to market requires corresponding pre-investment, along with respective metering and charging models in support of the corresponding business model. Therefore, making the characteristics flexible to artefact developers is not possible as it would be very hard to make the corresponding costs flexible and by that very hard to predict.

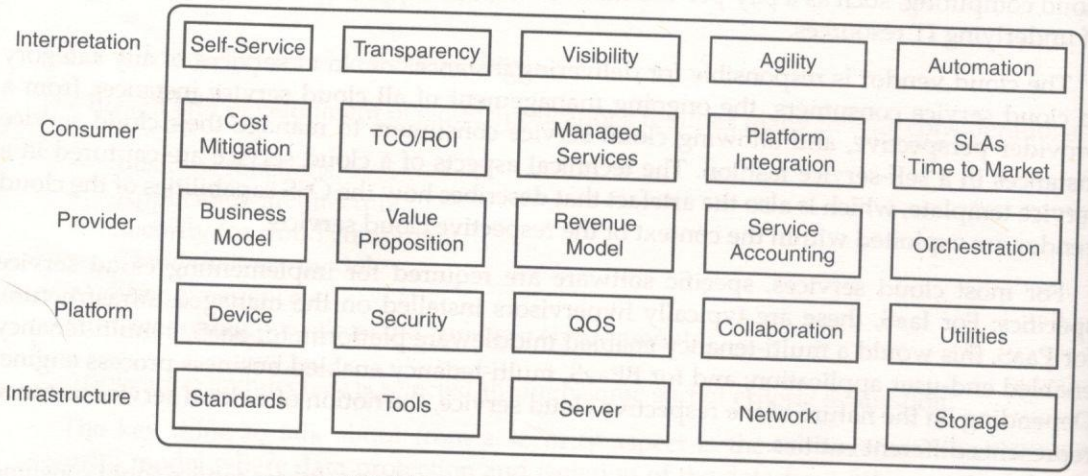


FIGURE 4.1 Cloud ecosystem.

This illustrates that defining and delivering a cloud service requires nailing down all corresponding functional and non-functional requirements. The artefacts developed on top of an ecosystem-focused cloud service have then only very minimal room to change how these functional and non-functional requirements are addressed (Figure 4.1). Note that this is not to be viewed as something negative, but rather as something very positive from an ecosystem perspective – it is a core value proposition of ecosystem-focused cloud services to provide pretty strict guidelines with respect to how they can be exploited as this is the main factor driving a reduction in cost of artefact development. The easier it is to develop artefacts for such a cloud service, the more likely the cloud service is successful.

As a summary, it is important to note that there is a difference between developing cloud services as a very conscious technical and business decision and developing artefacts on top of ecosystem focused cloud services prescribing the boundaries for how these artefacts can run.

Note that sometimes the concept of a 'Cloud Service' is also referred to as a 'Cloud Service Product'. The cloud's capability to have multiple environments to deploy the application is a

major advantage because it can be a mix and match condition that best fits the business function and the application. This means that the organization is not tied to one solution for cloud computing but rather multiple solutions. So, depending on the business needs, the application, and what the application has to offer, cloud requirements will help to select the proper cloud environments.

Cloud-based environments come handy especially when used to develop, test, and run your application for the following reasons:

- Available in your own private cloud environment or on the public cloud.
- Rapid access to a configurable development and test environment to speed time to market.
- Self-service Web portal for enterprise account management and provisioning in minutes.
- Pay-as-you-go pricing, with the choice of preferred pricing through reserved capacity packages.
- Security-rich environment designed to protect your systems and data.
- Access to a rich catalogue of software images for improved flexibility and rapid provisioning.
- Rapid provisioning and faster time to value.

4.3 CLOUD BUSINESS PROCESS MANAGEMENT

Business Process Management (BPM) governs an organization's cross-functional, customer-focussed, end-to-end *core business processes*. It achieves strategic business objectives by directing the deployment of resources from across the organization into efficient processes that create customer value. Its focus is on driving overall bottom line success by *integrating verticals and optimizing core work* (for example, order-to-cash, integrated product development, integrated supply chain). This is what differentiates BPM from traditional functional management disciplines.

In addition, intrinsic to BPM is the principle of 'continuous improvement', perpetually increasing value-generation and sustaining market competitiveness (or dominance) of the organization. It clearly defines and aligns operations, organizations, and information technology. The cloud environment can help in the following ways:

- **Integration of core business:**
  - Holistic.
  - Crosses organizational functions and boundaries (height and breadth).
  - Includes business and technology.
- **Value-focused efficiency:**
  - Customer-centric perspective.
  - Bottom-line success.
  - Speed at which ROI is delivered.
  - Performance measurement.
- **Continuous:**
  - This is based on longer period of intervals pertaining to cloud business.
  - Continual improvement.

- **Cultural:**
  - Cultural considerations of the organization and geographical area kept in mind at the time of due diligence of the requirement.

4.3.1 *Identifying BPM Opportunities*

This section discusses the opportunities required for successful cloud business process management and the characteristics of cloud deployment offerings. The answers to the following questions can help you identify cloud opportunities better:

- Are the strategic value proposition and capabilities defined for your organization?
- How does your overall strategy drive the design and execution of your business processes? Is there a traceability of execution to goals?
- How do you manage your core business processes?
- What are your current process initiatives?
- What are your current process governance facilities?
- Are your existing organizational structures aligned to enable efficient process operations?
- How do your customers measure and assess the performance of your processes?
- How does your process performance compare to your competitors?
- How effectively does your current technology (information, systems, tools, machines, etc.) enable the enterprise's core business processes?
- What risks and challenges does your current technology present for current and future process capabilities?
- What products does your organization have? What type of products?
- What are the notable pieces of your IT portfolio?
- Has your organization adopted SOA?
- How are processes currently modelled in your organization? What's included in the model?
- What design/development tools are currently used in the organization?
- What testing tools are currently used by the organization? What are the strengths/weaknesses of the tools?
- Please describe the different business processes that are automated in the organization?

Cloud application development offerings provide:

- Cloud application reference architecture.
- Unmatched experience developing high-performing, secure applications across a wide range of technologies of the cloud vendor.
- Unmatched application security expertise.
- Leadership in cloud related technologies – multi-tenancy, virtualization, pervasive computing.
- Significant expertise with cloud business models.
- Ability to integrate a portfolio of related cloud services.

4.3.2 *Cloud Technical Strategy*

This section gives the technical strategy on how the cloud customers can enable the cloud deployment.

Cloud services enable our cloud users to build middleware clouds in their datacenter and utilize public clouds, where it make sense by providing the following cloud-enabled middleware services:

- Infrastructure Services.
- Platform Services.
- Application Services.

Cloud strategy enables our organizations to do the following:

- Build middleware clouds in their datacenter.
- Utilize public clouds where it makes sense.

It does so by providing support in the following areas:

- **Cloud-enabled middleware services:**
  - Infrastructure Services.
  - Platform Services.
  - Application Services.
  - Serving the on premise and public clouds.

So, what does it mean to develop an application service for the cloud? For one, it means product features development similar to on-premise software. This means:

- **Enabling the software for cloud essentially implies:**
  - Support for collaborative multi-tenancy.
  - Self-service registration.
  - Managing customers and their entitlements.
  - Single sign on.
  - Additional security concerns.
- **Integration with the datacenter:**
  - Firewalls, reverse proxy configurations.
  - Fully qualified domain names, certificates.
  - Management of services, patch procedures.
  - Isolation, recovery, backup issues.

4.3.3 *Cloud Use Cases*

Infrastructure as a Service (IaaS) or Test/Development

Problem: Development teams require unpredictable amounts of infrastructure to get their job done. In majority of the cases, getting all of these resources in place before they are required in the development cycle can be quite a challenge. Purchasing the hardware consumes project budget and procurement of it is often quite slow. Static development and testing resources require manual re-provisioning in order to re-purpose resources for use, or new resources need to be purchased to meet demand. In cases where project timelines are short, project managers often choose not to set up much of an environment because it depletes their budget or

jeopardizes the project's delivery schedule. Actual usage of the system(s) can be quite short in terms of absolute time. Different types of projects need different kinds of development components (like SQL server, SharePoint, BizTalk, etc.) depending on the architecture of the solution. Besides the testing environment, the team usually needs a system that looks quite similar to the production environment in order to perform simulation, stress and load tests, for example, or to delivery end-user training, etc.

Solution: Companies can create standardized service catalogue items for common infrastructure requirements and enable development and test teams to access infrastructure in a self-service model (IaaS). Overall control over the process is maintained with business policies around quotas, reservations, reclamation, and standardized offerings.

Standardized Development Platforms/Middleware (PaaSEnable)

Problem: Developers are often not concerned about the impact of their code on Operations. They deliver their code without involving Operations into architectural decisions or code reviews. Enterprise architects recognize the large costs associated with non-standard development platforms. To simplify the ongoing maintenance and streamline development operations, many companies are creating corporate standards around development stacks that include middleware/applications. However, most companies are hesitant to an external PaaS offering due to the constraints of having to rewrite their internal applications to fit the external PaaS API sets.

Solution: Companies can create standardized development platform definitions for use by development teams to standardize and streamline their efforts. This improves corporate IT productivity by helping them build a 'private cloud' that provides a common foundation for building custom applications which run securely behind their own firewall.

Application Cloud

Problem: Companies want to move beyond self-service for infrastructure and provide application owners the ability to define, instantiate, and manage complex multi-tier applications. This includes configuring the application for production usage and monitoring performance within the application for SLA optimization.

Solution: End users can access complete application definitions and manage them according to their quotas and preferences defined by cloud administrators. Applications in production can be monitored across multiple factors and automatically scaled up and down according to business policies.

Software-as-a-Service (SaaS) to End Customers

Problem: Many companies or ISVs want to deliver their applications to end users as a service. However, creating a multi-tenant SaaS offering requires substantial development to support security, performance, and scalability needs. Due to these high costs, companies cannot offer new services based upon existing applications.

Solution: Companies can provision a unique application instance per customer with private cloud automation capabilities. Environments are provisioned according to business policies and unique SLAs can be delivered per customer according to the business arrangements.

The cloud engine should provide the automation to provision on-demand complex application and configuration environments required along with the dynamic application scaling. It should also deliver unique business policy selections including custom placement and high availability across multiple datacenters.

4.4 CLOUD SERVICE MANAGEMENT

A service management system provides the visibility, control, and automation needed for efficient cloud delivery in both public and private implementations:

- **Simplify user interaction with IT:**
  - User-friendly self-service interface accelerates time to value.
  - Service catalogue enables standards which drive consistent service delivery.
- **Enable policies to lower cost with provisioning:**
  - Automated provisioning and de-provisioning speeds service delivery.
  - Provisioning policies allow release and reuse of assets.
- **Increase system administrator productivity:**
  - Move from management silos to a service management system.

The emergence of cloud deployments is prompting enterprises to either assemble in-house teams to manage specialized cloud service providers or look to third-party cloud brokers chiefly due to the following reasons:

- Every service-oriented approach needs a mechanism to enable discovery and end-point resolution.
- Registry/repository technology provides this where service delivery is inside the firewall.
- Cloud services delivered across firewalls need something similar – a third party that serves as a 'service broker'.

Leveraging service brokers will probably become a critical success factor in cloud computing as cloud services multiply and expand faster than the ability of cloud consumers to manage or govern them. The growth of service brokerage businesses will increase the ability of cloud consumers to use services in a trustworthy manner. Cloud service providers are expected to begin to partner with cloud brokerages to ensure that they can deliver the services they promote. These cloud intermediaries will help companies choose the right platform, deploy apps across multiple clouds and perhaps even provide cloud arbitrage services that allow end-users to shift between platforms to capture the best pricing.

There can be three categories of opportunities for cloud brokers:

- **Cloud service intermediation:** Building services atop an existing cloud platform, such as additional security or management capabilities.
- **Cloud aggregation:** Deploying customer services over multiple cloud platforms.
- **Cloud service arbitrage:** Supplying flexibility and 'opportunistic choices' – and fostering competition between clouds.

It will be similar to cloud services under one umbrella except that the services being aggregated won't be fixed. This flexibility will be important while doing chores such as providing multiple e-mail services through one service provider or providing a credit-scoring service that checks multiple scoring agencies and then selects the best score.

The ability to federate an application across multiple clouds will become important – if one service goes down, another can be started – and the service broker will just simplify it. To help federate the clouds, a 'storefront' (Apps.gov) site can be created with services that are pre-screened to meet government procurement guidelines. The new site can be expected to cut red tape and make it easier for government agencies to quickly deploy the latest technology.

4.4.1 *Key Cloud Solution Characteristics*

The essential cloud orchestrator and engine key characteristic capabilities are:

- **Scalability:** Cloud orchestrator should maintain an index of the resources that are acquired from the hypervisor, giving the master a low overhead and enabling it to scale across tens of thousands of machines across multiple geographies.
- **High Availability:** Cloud orchestrator should play for the master node to support 'Active–Passive' as well as 'Active–Active' scenarios for availability and Disaster Recovery (DR). Cloud orchestrator should also monitor individual physical server for availability and in case of a physical resource server failure, should restart the VM on another running server to meet the requirements.
- **Application Lifecycle:** Cloud orchestrator should offer complete application lifecycle support from the creation of infrastructure to installation, configuration, and launching an application to deletion or expiration. This allows applications to be instantiated, removed, or flexed very quickly to respond to real-time demand for those applications.
- **Multi-tenancy/Role-based Administration:**
  - Cloud orchestrator should support multi-tenant capability with specific user permissions. Cloud orchestrator should have multiple personas which are like Cloud Admin, Account Owner, and User. Application definitions are only 'published' to specific users. The application owner or administrator logs in with his/her credentials and can view (User) the application VMs that have been allocated to him/her and do admin operations (Account Owner) on those VMs if needed.
  - Role-based administration allows fine-grained control of what each person can or cannot do in terms of cloud orchestrator features.
- **Policies:** Cloud orchestrator should provide rich set of policies that can be enabled. These policies can be modified or new ones can be created to take effect at the global level on applications, VMs, hosts, etc. These policies can also be embedded in the service definitions to take effect automatically depending on metric threshold. For example, a policy can be created to allow an application to flex up to 10 VMs during high load or demand and to be reduced to only 2 running VMs during low load or demand. This frees up resources that can be used by other applications that are experiencing high load.
- **Alarms:** Cloud orchestrator should provide pre-defined alarms that can be set at the global level for applications, VMs, hosts, etc. These alarms can be used to notify individual users or application owners regarding the application thresholds being reached.

For example, an alert can be sent if the response time of an application is below an SLA but there are no more resources for the application to flex-up.

- **Application Awareness and Policy-based Allocation:** Cloud orchestrator should be aware of application requirements and optimizes the placement of the application accordingly, e.g. placing the VMs running the application close to each other to reduce latency. Cloud orchestrator should support the major application servers.
- **Resource Awareness and Policy-based Allocation:** Cloud orchestrator should optimize the usage of the cloud infrastructure through intelligent resource allocation policies and allows load balancing of the VMs
- **Elasticity Based on Performance (Flex-up/Flex-down):** Cloud orchestrator should provide out-of-box functionality to flex-up or flex-down an application instance or resource based on performance metrics.
- **Reporting and Accounting:** Cloud orchestrator should provide metering and billing reports on resource allocation and actual usage. Additionally, this data can be used to create reports on inventory capacity and consumption. This allows the different business owners to create reports on how much or how little the application is used, and administrators can then adjust the resources allocated to each application accordingly.
- **Self-Service Portal:** Cloud orchestrator should enable a self-service portal for application owners. Application owners can request machines or entire multi-machine application environment, monitor, and control them through this portal. It should drive the workflows necessary to create the environment, and provide run-time environment management in order to support application elasticity. For example, the owner of the auditing application may request more resources for his application during a busy period.

4.5 ON-PREMISE CLOUD ORCHESTRATION AND PROVISIONING ENGINE

On-Premise Cloud Orchestration and Provisioning Engine can be a bundled offering that includes hardware, software, and the services one needs to get started with cloud computing. It should include all the elements in a services ecosystem. It must have a self-service portal, it should include the automation, and it should track and control all of resources.

The other objective is to completely integrate and include a service and then, on top of that, users can add additional services to do integration or other types of cloud work. It can be a pre-packaged private cloud offering that can bring together the hardware, software, and services needed to establish a private cloud to accelerate selling efforts and effectiveness.

Cloud Orchestration and Provisioning Engine should be designed from client cloud implementation experiences and integrated with the service management software system with servers, storage, and services to enable aprivate cloud in IT environment. This will help to remove the guess work of establishing a private cloud by pre-installing and configuring the necessary software on the hardware and leveraging services for customization to your environment. All that is required is to install your applications and start leveraging the benefits of cloud computing, such as virtualization, flexibility, scalability, and a self-service portal for provisioning new services.

Cloud Orchestration and Provisioning Engine should provide an alternative to traditional IT infrastructure for IT executives seeking to enhance delivery of services and to transform the datacenter into a cost-effective Dynamic Infrastructure. Cloud Orchestration and Provisioning Engine ecosystem should be 'Built for performance' and should be based on architectures and configurations required by specific workloads. It should enable the datacenter to accelerate the creation of services for a variety of workloads with a high degree of flexibility, reliability, and resource optimization.

4.5.1 *Benefits/Value Proposition*

Powers faster time to innovation, lowers cost per unit of innovation

- **Innovation:** It should dramatically improve business value and IT's effect on time-to-market by enabling the business workloads to rapidly and accurately be deployed when and where they are needed.
- **Decrease Operational Expenses:** It must gain productivity increases in IT labour costs through automation. Maximize capital usage and reduce added capital expense.
- **Reduce Complexity and Risk:** With automation and standardization, the human error-factor should be minimized.

4.5.2 *Cloud Orchestration and Provisioning Requirement Analysis*

In order to understand the Cloud Orchestration and Provisioning Engine requirements, we need to understand the test and development requirements of the cloud deployment. Therefore, we should be sure about the automation of the testing and development cycles to reduce the deployment time. In order to do so, we can initiate the process of discussion and try to talk to cloud customer about the opportunity for the deployment of the Cloud Orchestration and Provisioning Engine. After this, we can set the boundaries of the environment. About 30–50 percent of any given IT environment is devoted to test/development purpose. It can take developers days, weeks, or even months to procure and configure appropriate hardware, networking, management software, storage, with which they can test. With the Orchestration and Provisioning Engine, a developer can log into a self-service portal, select resources required and timeframe, select an image to provision from the service catalogue, and be ready to go in hours as opposed to months. Customer datacenters support hundreds or thousands of distinct composite applications representing business workloads. They leverage multiple types of servers, storage, networks, middleware, and operating systems. About 70 percent of the IT datacenter expense is spent assembling and re-assembling existing infrastructure, leaving fewer resources for innovation.

Cloud solutions are based on its service management solutions. The first solution in a series of Cloud Orchestration and Provisioning Engine solutions that we should consider should be workload-specific deployments. It will be a great entry point for users who want to get into cloud computing as you can roll a Cloud Orchestration and Provisioning Engine into their environment without effecting anything else in the environment, and use it as your initial pilot project to start running cloud services.

Cloud Orchestration and Provisioning Engine in any IT environment represents an alternative entry point into cloud computing. So, in some cases, the users want to turn their existing

environment into a cloud. In that case, we go into the datacenter, install a cloud management platform, and assign the existing resources that are there to the cloud. One scenario for this is that the organization already has plenty of equipment that they are just not using efficiently, and they see a lot of benefit from turning their existing investment into a cloud.

Cloud Orchestration and Provisioning Engine can be a great door opener and a great way to jump-start your cloud services. It can be bundled with the hardware, software, and services that you need to quickly getup and running; you can use this as a seed-and-grow model.

Entry points:

- Turn existing environment into a cloud:
  - Install cloud management platform and assign existing resources to the cloud.
  - Scenario – you already have enough equipment or Cloud Orchestration and Provisioning Engine offering is not the right platform.
- Jump start your cloud
  - Hardware + Software + Services required for a quick start up.
  - Can use a 'seed and grow' model – start with a Cloud Orchestration and Provisioning Engine and then addmore.

So what solution should you use? Let's look at some scenarios. So when can you opt for Cloud Orchestration and Provisioning Engine? This should be when you wish to get started with cloud computing and see the advantage of having a cloud management platform bundled along with the resources. This is the most rapid way to get a cloud platform up and running. In a scenario where you want to transition your existing resources into a cloud, you use this service. Also, a user who uses Cloud Orchestration and Provisioning Engine will benefit when cloud vendor can help them plan, design, and implement services, whether they want to implement infrastructure as a service, platform as a service, or turn it into a production cloud.

So, in some cases as you can see, it's appropriate to opt for Cloud Orchestration and Provisioning Engine and the business development and test services, because even if you just start with provisioning engine, you may eventually want to build out your service catalogue, integrate it into your directory, integrate it into cloud vendor or third-party service management products, and extend the capability of the platform.

4.5.3 *Cloud Infrastructure Security*

The security aspect of the cloud infrastructure goes side by side with Service Oriented Architecture (SOA) security. We can introduce it as a layered approach. At the top we can see the service layer with the run-time secure virtualized environment. This is available as the distributed service environment. These services include administrative and security aspects across different clouds as well as within a single cloud. This should gel with Web services stack and it is important that we should bind the internal resources with the other cloud services to offer better hybrid cloud services for successful cloud deployments. We will discuss more about this in Chapter 9.

One of the key aspects of SOA is the ability to easily integrate different services from different providers. Cloud computing is pushing this model one step further than most enterprise SOA

environments, since a cloud sometimes supports a very large number of tenants, services and standards. This support is provided in a highly dynamic and agile fashion, and under very complex trust relationships. In particular, a cloud SOA sometimes supports a large and open user population, and it cannot assume a pre-established relationship between cloud provider and subscriber.

The Secure Virtualized Runtime layer on the bottom is a virtualized system that runs the processes that provide access to data on the data stores. This run-time differs from classic run-time systems in that it operates on virtual machine images rather than on individual applications. It provides security services such as antivirus, introspection, and externalized security services around virtual images. While the foundations of Secure Virtualized Runtime predate SOA security and are built on decades of experience with mainframe architectures, the development of Secure Virtualized Runtime is still very much in flux. Cloud vendors continuously invest in research and development of stronger isolation at all levels of the network, server, hypervisor, process, and storage infrastructure to support massive multi-tenancy.

Cloud Orchestration and Provisioning Engine Integrated service management is offered with network, servers, storage, services, and financing as an integrated offering for client test platforms.

- **Improved time to value:** Quickly deliver a cloud using a preloaded and integrated system.
- **Improved innovation:** Dramatically improve business value and IT's effect on time-to-market by delivering services faster.
- **Decrease capital expenses:** Maximize capital usage and reduce added capital expense.
- **Reduce complexity and risk:** With automation and standardization the human error factor is minimized.
- **'Fit for purpose':** Based on architectures required by specific workloads.
- **Self-contained:** Service management, software, hardware, storage, networking.
- **Modular:** Automatically expandable and scalable.
- **Virtualized:** End-to-end across server, network, and storage.
- **Self-service:** Ease of consumption.
- **'Lights-out':** Zero touch automated operations.

Cloud Orchestration and Provisioning Engine is offered as a services engagement which can build a solution to a client's needs, including creation of custom virtual images for dispensing. Summary points are given below

- **Drastically reduce set-up and configuration time:**
  - New environments in minutes!
- **Reduce risk by codifying infrastructure:**
  - Freeze-dry best practices for repeated, consistent deployments.
- **Security throughout the entire lifecycle.**
- **Simplify maintenance and management:**
  - Flexibly manage and update the components of your patterns.
  - Ensure consistency in versions across development, test, production.
- **Spend less time administering, more time developing new solutions.**

4.6 COMPUTING ON DEMAND (CoD)

On-demand computing is the need of the hour. It is very essential even in a supercomputing environment. On-demand computing can be implemented using various virtualization techniques. Cloud gives you an option to leverage the computing infrastructure without actually buying the hardware. This helps you to transfer the workload if your resources are not able to support it, and at other times, lets others utilize your resources that are lying idle. In this way this makes it possible to use the resources in most efficient ways. It may be possible that there can be many spikes that can come for the utilization but cloud experts can make it smooth.

The uniquely rich set of features that on-demand computing can offer enables service seeker to deploy a true utility. The platform allows users to:

- **Align cost with utilization** so that users can scale costs down as well as up. This allows a workload to start with minimal upfront costs and scale as the demand grows without paying a penalty to increase capacity. Additionally users can benefit from not incurring the disruption to move to a larger machine.
- **Increase end-users availability significantly.** As workload can be moved dynamically, it is possible to move workload from one server to another without interruption so remedial work can be carried out if server down time is required.
- **Balance workload dynamically across multiple servers** without taking applications offline. Using the workload mobility features customer can align their costs by ensuring that workload is deployed in such a way as to optimise systems resource.
- **React to short-term resource requirements** almost instantly. If a workload has to be deployed at short notice, a virtual machine (VM) can be created on the server and resources allocated instantaneously using the dynamic capacity model.
- **Reduce the physical foot print** in the datacenter. Consolidation of workload on to a smaller number of servers will improve space, power, and cooling metrics.
- **Confidently increase system utilization** to over 75 percent without fear of degrading performance for end-users.
- **Develop a simple charging model that reflects usage** for end users as the service delivery culture continues to mature.
- **Double the workload delivered** in the power and cooling envelope.

4.6.1 *Pre-Provisioning*

For the on-demand computing requirement pre-provisioning is the viable option as it helps organization meet the requirement of the dynamic datacenter requirements. Organizations would like to reduce the time they take to commission servers when a new workload is to be deployed.

This approach is ideal when:

- The sizing and capacity planning is fully understood.
- The workload is fairly constant, ensuring good utilization levels are achieved.
- There are business reasons that require the physical separation of workload.
- Workload can be scaled horizontally.

4.6.2 On-Demand CPU/Memory/VM Resources

In the dynamic environment it is important to track the requirement of CPU, Memory, and VMs. It is based on the common pool concept where resources are allocated and de-allocated as the requirement is over. This approach is ideal when:

- Workloads are trending upwards so investment can be aligned with utilization.
- Peeks in workload are longer term.
- Workload scales vertically.
- It is more economically advantageous to 'buy out' dynamic capacity.

4.6.3 Dynamic Capacity

Utility CoD is used to automatically provide additional processor capacity on a temporary basis within the shared processor pool. Usage is measured in processor-minute increments, and is reported via a Web interface or collection of report by cloud vendor engineer. Billing is based on the reported usage.

This approach is ideal when:

- The workload is very variable and multiple workloads can be hosted on a single machine so that the utilization can be levelled out.
- The workload has short periods where system utilization increases massively, but for the majority of the time it is not resource-intensive.
- Workloads can share a physical platform.
- The workload is designed to scale vertically only.
- Users want to dynamically balance workloads across servers.
- Users want to continue to run very small workloads without incurring the overheads associated with running a physical server to support it.

Benefits

uncapped - we can add on resources, deleted =

- Partition mobility, significantly reduced power/cooling footprint, donation of unused processor cycles of VMs with dedicated processors to uncapped partitions and at the same time guaranteed performance of these VMs.
- Very short deployment time (time-to-market optimized).
- Lowest possible cost for deployment of small workloads.
- Less management effort, for example, when using VMs.
- Most granular charging scheme, pay for the CPU and memory cycles actually used.
- Complete decommissioning of partitions; resources are available for other purposes.
- Flexible workload management, workloads can compensate for each other, thus reducing overall utilization.
- Ideal for environments with identical systems management, utilities for development and testing.

Limitations

- Short peaks, must not exceed certain limits, and needs to be monitored (via Web interface) to ensure best value is obtained.
- Utility CoD provides processor resources only to the uncapped partitions.

However, one of the most important advantages of the Dynamic Capacity model is to be flexible, that is, to allow the switching of CPU capacity on and off as needed, which can reduce costs significantly. This not reflected in this calculation as it requires input from application owners.

4.6.4 *Cloud Platform Characteristics Based on CoD*

This section discusses cloud platform characteristics on the basis of low-end, on-demand, and dynamic-capacity–based servers.

Low-End Servers

- Physical segregation of servers.
- High administration cost due to management of more physical servers.
- Limited and complex scalability, process – maximum 8 processors per server – slower turn-around time for server deployment.
- Longer lead time for server deployment from ordering of servers to setting up of infrastructure.
- Not ideal for short product lifecycle application due to fixed cost expenditure for hardware.
- Wastage of hardware resources for applications that reacts to volatile market.
- Unable to share resources between applications.
- Wastage of un-used processing cycles if the application does not fully utilized the resources.
- No hardware/application interdependency forcing down time on application owners.
- No capacity on demand capability. Downtime is required for adding new hardware.
- Low price per CPU cycle purchased but higher cost per CPU cycle actually used.

On-Demand Platform

- Physical or logical segregation of servers or partitions implementation.
- Lower administration cost due to less physical servers' management.
- Can cater for quick turn-around time for new application deployment or increase capacity due to business requirements.
- Enhanced time to market for new product launch with immediate availability of CPU/memory capacity.
- Not ideal for short product life cycle application due to fixed cost expenditure for hardware.
- Wastage of hardware resources for applications which reacts to volatile market.
- Able to share I/O, CPU, and memory resources between applications.
- Able to take advantage of un-used CPU/memory if dynamic VM reallocation or share pool methodology is implemented.
- To provide an environment in which there are no hardware/application interdependencies forcing down time on application owners care full capacity planning and management is required.
- Capacity on demand capability – no downtime is required if COD CPU/memory is sufficient.
- Higher price per CPU cycle, lower cost per CPU cycle actually used.

Dynamic Capacity Platform

- Choose virtual machine or workload virtual machine implementation for application consolidation.
- Lower administration cost due to less physical and logical servers management.
- Can cater for quick turn-around time for new application deployment.
- Enhanced time to market for new product launch with immediate availability of infrastructure and setup.
- Ability to scale up and down which will be ideal for application with short product life cycle.
- Able to cater for applications which react to volatile market, i.e., scaling up and down capacity.
- Able to share I/O, CPU, and memory resources between applications.
- Able to take advantage of un-used processing cycle of other applications.
- No hardware/application interdependency forcing down time on application owners as workload can be dynamically moved to facilitate maintenance, etc.
- Capacity on-demand capability. No downtime is required as the machine is fully configured.
- Higher price per CPU cycle, lower cost per CPU cycle actually used. Average price due to the ability to optimize utilization and rapidly deploy workload.

4.7 CLOUDSOURCING

Today we are living in the era of optimizing the hardware resources and moving towards the large enterprise day by day so cloud computing is becoming the ingredient part of infrastructure deployments. Now you may not need only cloud computing, you may need the entire consulting, implement and management solutions.

The new wave that is igniting the cloud deployments in the service industry as a new trend is Cloudsourcing – outsourcing the end to end solution using cloud methodology using public cloud, infrastructure and platforms. This will be more planned approach as it will comprise the whole service cycle of outsourcing business with cloud principles with the help of strategized connected cloud platforms that will match the overall enterprise requirements.

This includes the whole cloud implementation, IT business consulting, integration, and configuration of the business. This will give the option through which we can enjoy the benefits of service industry with the benefits of the clouds that gives the innovative approach of paying the resources over subscription.

Real deployment of the Cloudsourcing will requires the business model with the impact of cloud customer and cloud vendor requirements.

With respect to cloud customers, it is important to note there is no control on the infrastructure layout of the cloud deployments. Even there is no control over the place from where the data services are offered from the cloud vendor. It is also to known that cloud customer don't have think about the operational staff for the deployments.

Thus, Cloudsourcing will be playing the vital role in the next generation of cloud implementation. With the availability of new open source tools it is like icing on the cake, integrated with partner cloud solutions, platforms and infrastructure. Also the new charging models like the services on both a project and subscription basis will give new wave to deploy and adopt the cloud sourcing models.

This will help to customize application on cloud infrastructure. This will be primarily being offered as a public cloud and all these offering will be available as managed services. These services will be prototype based that is developed internally on the product and working applications. Therefore, it will give a good chance to use the intellectual property for developing different business vertical solution easily.

4.8 SUMMARY

In this chapter, we pointed out the main features of Cloud Orchestration and Provisioning Engine, BPM clouds, cloud sourcing, and requirements of service management. Next chapter will discuss about different types of cloud offerings.

CHAPTER 5

CLOUD OFFERINGS

Introduction

Information Storage, Retrieval, Archive, and Protection

Cloud Analytics

Testing Under Cloud

Information Security

Virtual Desktop Infrastructure

Storage Cloud

Summary

5.1 INTRODUCTION

Information is pouring in faster than we can make sense of it. It is being authored by billions of people and flowing from a trillion intelligent devices, sensors, and instrumented objects. With 80 percent of new data growth existing as unstructured content from music files, to 3D images, to medical records, to e-mail keystrokes, and more, the challenge is trying to pull it all together and make it useful.

Until now, organizations could not fully or quickly synthesize and interpret all the information out there – they had to make decisions based largely on instinct. But now, there are mechanisms that can capture, organize, and process all the data scattered throughout an organization, and turn it into actual intelligence. This enables organizations to make better business decisions.

5.2 INFORMATION STORAGE, RETRIEVAL, ARCHIVE, AND PROTECTION

Organizations process, manage, move, protect, and archive various business data according to unique characteristics such as age, usage patterns, compliance and archiving policies, security and disaster protection rules, and value. Information Lifecycle Management (ILM) is a growing set of recommended practices and technologies to manage data more efficiently and effectively. Lifecycle management becomes even more important for cloud deployments as we would be sharing the data services between the cloud vendor and subscribers.

ILM is not the latest data storage, retrieval, and protection solution; piece of hardware; or some software, but rather an approach to assess and manage information across the enterprise. ILM is based on how data is used and how readily available it must be to the people who use it. It is focused on managing and storing data according to its value to business operations at any given point in time. It is concerned with placement of data on the appropriate level of storage with the appropriate retention and retrieval policy. In response to these challenges, organizations are defining specific objectives to support and improve their information management:

- **Cost reduction:**
  - Controlling demand for storage.
  - Reducing hardware/software costs.
  - Reducing storage personnel costs.
- **Better system performance and personnel productivity (i.e., improve efficiency):**
  - Doing the storage activities 'right'.
  - Improving the current people, processes, and technologies being utilized to deliver storage services to the business.
  - Defining and enforcing policies to manage the lifecycle of data.
- **Increased effectiveness:**
  - Doing the 'right' storage activities.
  - Defining and implementing the appropriate storage strategy to address current and future business requirements.

They are also coming up with new ways to generate, enhance, and sustain higher savings. These include:

- **Activities for gaining initial savings:**
  - Reduce the amount of used storage as a result of initial clean-up.
  - Validate SAN requirements and reclaim used switches and switch ports.
  - Validate data replication requirements in order to reclaim used storage space and offset future growth requirements.
  - Develop and document information classification.
  - Develop and document classes of service.
  - Design and implement the tiered storage architecture.
  - Migrate existing information to lower cost storage using a tiered storage architecture.
- **Activities for maximizing savings:**
  - Reconfigure the current storage environment effectively improving the available to raw utilization.
  - Reclaim available storage that has been over allocated.
  - Enhance the information classification, classes of service, and tiered storage architecture.
- **Activities for sustaining savings:**
  - Develop a storage architecture governance model.
  - Implement changes to existing storage management processes like capacity planning and provisioning that to effectively improve capacity utilization on an ongoing basis.

While designing a target storage environment, the estimated financial impact is calculated based on the following key cost components:

- **Operating cost categories:**
  - **Personnel:** Storage support and contractors.
  - **Facilities:** Current floor space consumed by storage, telecommunication charges attributed to storage and tape vaulting services.
  - **Storage hardware maintenance:** Existing maintenance and incremental maintenance resulting from growth.
  - **Storage software maintenance:** Existing maintenance and incremental maintenance resulting from growth.
  - **Outages:** Cost avoidance associated with the reduction in unplanned outages.
- **Investment cost categories:**
  - **New hardware required:** Typically includes disk, tape, and array cost but not the incremental cost of adding SAN fabric. Investment is either upfront or over a period of time if the client leases equipment.
  - **New software required:** New storage software required to support the target environment.
  - **Hardware refresh:** Investment required to refresh the existing hardware is often considered in the base case.
  - **Transition services:** Incremental cost required to migrate the current environment to the future environment. Not typically estimated until the scope of the third-party implementation services has been defined.

* hardware that connect the workstation & server to storage devices.

When more than 90 percent of the data stored on hard disks is not actively accessed by users or applications, it is obviously ripe for more intelligent management and migration to less expensive storage. But the savings can go significantly beyond disk acquisition costs and annual hardware maintenance costs.

Some points in Information Management:

- **Data:** Discrete element, reasoning, discussion, or calculation of content created through the interactions between applications or interactions between computing devices.
- **Information:** Organized and structured collection of data.
- **Information Lifecycle Management:** The policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost-effective IT infrastructure from the time information is conceived through its final disposition.
- **Information Taxonomy:** Data described in the context of business process requirements and lifecycle characteristics.
- **Information Classes:** Groups of information taxonomies with associated business value that provide the basis for storage management and service delivery.
- **Value-driven Data Placement:** An event correlation framework that 'senses' when the value of data changes and based on business policies 'responds' by moving data to the appropriate storage tier.
- **Storage Process:** A documented set of storage-related tasks and activities required to support a storage infrastructure.
- **Storage Service:** Capabilities provided to a customer base designed to meet their business requirements, wants, and needs that is enabled by a storage infrastructure.
- **Enterprise Class of Service (COS):** A common set of storage services that are delivered to meet a corresponding set of storage requirements based upon key information management characteristics and the features, functions, capabilities, processes, and governance required to deliver the required enterprise storage services.
- **Storage Tier:** A subset as a set of storage devices that are identified to store and/or maintain information for a predefined period of time based on key information management characteristics such as performance, configuration, residency, retention, value, etc.
- **Tiered Storage Infrastructure:** An organized collection of Storage Tiers reflecting the flow of all information managed in the enterprise storage architecture.
- **Utility-based Service Delivery:** The 'just-in-time' delivery of standardized storage processes, management, and infrastructure, as a measurable service, on a 'pay-as-you-go' basis.

5.3 CLOUD ANALYTICS

Cloud analytics is the new offering in the new era of cloud computing. This will help in the consulting domain and will ensure better results. It provides users with a better forecasting technique to analyze and optimize the service lines and provide a higher level of accuracy. Cloud analytics can help them apply analytics principles and best practices to analyze the different business consequences and achieve newer levels of optimization (Figure 5.1). This can combine complex analytics with the newer software platforms and will lead towards the predictable business situation out of every business insight.

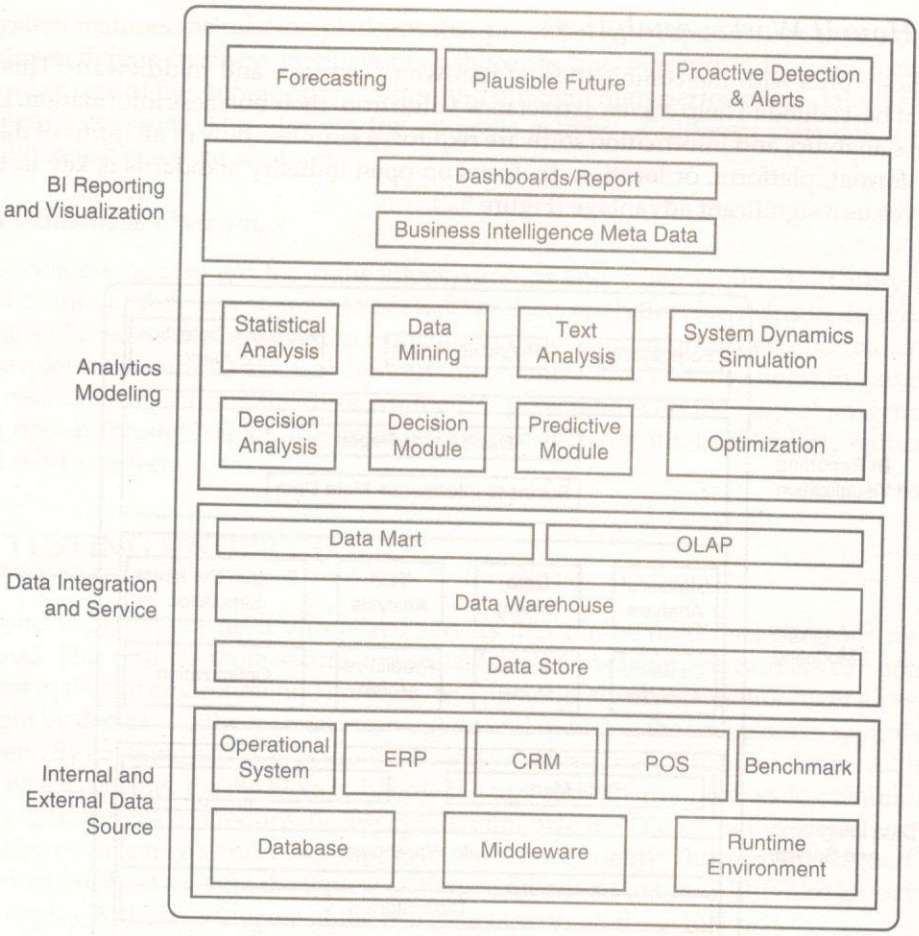


FIGURE 5.1 Cloud analytics.

5.3.1 *Cloud Business Analytics Competencies*

The cloud business analytics service line is supported by different types of competency areas. One of them is cloud business analytics strategy that helps clients achieve their business objectives faster, with less risk, and at a lower cost by improving how information is recognized and acted upon across the enterprise or within a business function. The next competency is business intelligence and performance management that helps increase performance by providing accurate and on-time data reporting. The next is analytics and optimization that provides different type of modelling techniques, deep computing and simulation techniques to check for different type of 'what if' analysis to increase performance. The other competency is enterprise information management that lets you apply different architecture related to data extraction, archival, retrieval, movement, and integration. Another competency that is required for the cloud analytics is the content management system that includes the different service architecture, technology architecture, and process related to capturing, storing, preserving, delivering, and managing the data. It also helps to provide access in the global environment and makes it easy to share data with stakeholders across the globe.

5.3.2 *How It Works: Analytics*

Analytics works with the combination of hardware, services, and middleware. This expertise makes it best suited to help clients extract new value from their business information. Delivering business analytics and information software requires a seamless flow of all forms of data regardless of format, platform, or location. Its focus on open industry standards is key to this effort, and gives us a significant advantage (Figure 5.2).

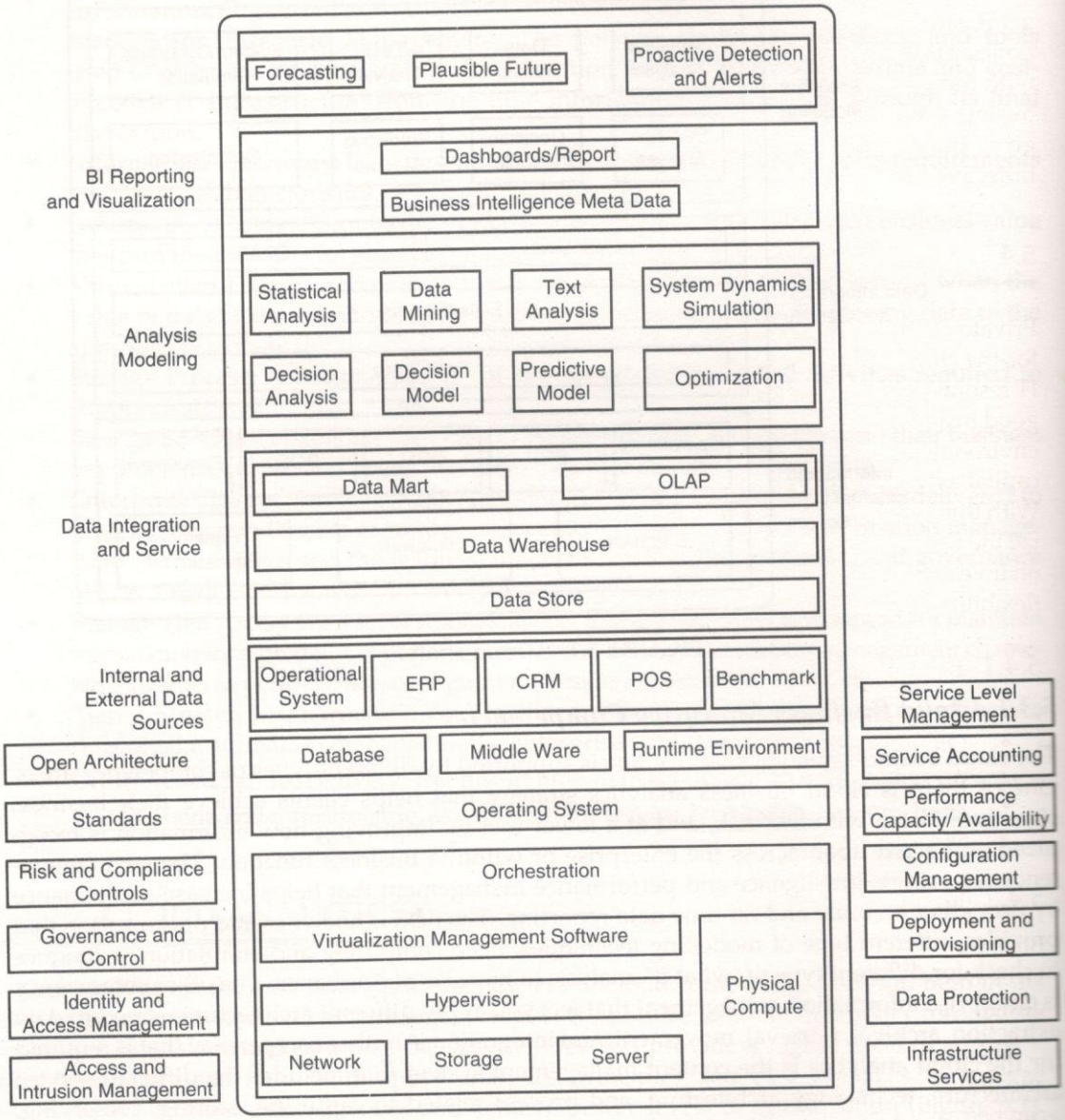


FIGURE 5.2 Cloud analytics business outcomes.

The system features include the platform that provides data reporting, analytics based on text, mining activities, business intelligence dashboards, and perceptive analytics techniques. This also takes care of the storage optimization and different high-performance data-warehouse management techniques. This also has the umbrella activity of different installation services and a highly reliable system platform.

Analytics Business Outcomes

Analytics systems help to get the right information as and when required, identify how to get it, and point out the right sources to get it. Therefore, analytics also helps in designing the policies faster based on the information available in the organization as decision-makers work with the exploration services available within the organization. This also helps in gauging the business results by measuring the different metrics generated with the help of analytics. This gives the option through which the organization can increase the profitability, reduce cycle time, and reduce defects.

5.4 TESTING UNDER CLOUD

Private cloud deployments need virtualized servers that can be used for testing the resources for the cloud. This ensures more secure and scalable solutions where consumers can access the IT resources in the test environment. Therefore, testing under the cloud environment gives a very good insight by decreasing the manual intervention and reducing the processes in typical testing environment. By enabling access to resources as and when required, it reduces the investment on capital as well as enables the business to handle the ups and downs of the testing requirements. With this, organizations can reduce the test cycles, minimize the IT cost, reduce defects, rationalize the testing environment, and hence, improve the service quality. This provides a good return on investment (ROI) on moving the typical testing environment to cloud. This also gives you the flexibility to play with the surrogate of the real system without the actual risk.

5.4.1 Benefits

- Cut capital and operational costs and not affect mission critical production applications.
- Offer new and innovative services to clients, and present an opportunity to speed cycles of innovation and improve solution quality.
- Facilitate a test environment based on request and provide request-based service for storage, network, and OS.

5.4.2 Value Proposition

Business test cloud delivers an integrated, flexible, and extensible approach to test resource services and management with rapid time to value. This is an end-to-end set of services to strategize, design, and build request-driven delivery of test resources in a cost-effective, efficient manner.

These test tools allow you to orchestrate and build your services and development projects and allow you to catalogue and organize all of the various software assets that you have. These can be provisioned by administrators, development team leads, or project team members that you give permission.

5.4.3 *The Biggest Benefitters*

The biggest problems for the finance heads are to reduce the operational and capital expense of the development and testing environment. Testing under cloud environment comes as a boon to implement this. This not only helps financially to reduce the burden on the company as well as reduce the cycle time for testing and development environment without buying the infrastructure. This environment provide test tools to synchronize and build the services for the project and helps arrange the project resources, assets and versions of deliverables. Also, permission can be set based on the roles to provide the access to the assets.

So why is a test cloud a great on-ramp to cloud computing for organizations that want to embrace cloud computing and are trying to figure out where to start? Well, the test and development environment is rarely optimized and is usually low hanging fruit in terms of getting ROI and benefit out of cloud computing. These are environments that have a high degree of dynamic change. So, it is a lot of build-up and teardown as you move from project to project that you test and develop. Typically, 30 to 50 percent of all servers in an IT environment are dedicated to test and development, which is a significant amount. A number of these servers are sitting there at very low levels of utilization, and in fact, sit idle for a period during the project life cycle. So, being able to better utilize those systems can be a huge benefit for the organization.

Many defects are introduced into test and development environments because of the fact that there is a high degree of manual configuration that typically goes on. Often, projects are backlogged because of limited access to test environments. So, if you do a lot of test and development, or even if you buy your products, you still have to integrate those in a test environment before they move them into production; this intends to be an excellent entry point into cloud computing.

It is important to take a holistic look at the problem you are trying to solve. Are you building a development cloud? Are you building a test cloud? What pre-existing service management or enterprise management tools exist, so that you can identify the starburst opportunities that come off business development and test cloud services or at the cloud strategy and planning consulting workshop service? Some of the things to think about include:

- What are the types of service management integrations that need to happen?
- Do they want to integrate with discovery?
- Do they want to integrate with the service desk?
- Do they want to create a help ticket whenever a provisioning step occurs?
- Do they want to create an asset every time a new virtualized environment is created?
- Do they want to do usage and accounting chargeback?
- Do they want a test optimization service?

This and many other possible opportunities exist, so make sure that this is part of the requirements discussion so that it can make its way into the design and implementation piece of the engagement.

Organizations engaged in development and test share a common challenge in executing fluid projects within rigid infrastructures. Development projects are initiated to meet specific 'needs', whether bringing new applications to market, updating or patching existing software, or developing in-house assets. There is no way to 'schedule' needs; identification and

application development can be as much art as science, with an intrinsic unpredictability that makes resource allocation very difficult within traditional infrastructures.

As needs arise erratically and projects slip within their individual timelines, there are inevitable overlaps and gaps in development resource requirements. Organizations are often faced with frustrating delays as projects wait idle for infrastructure availability. Alternatively, maintaining enough capacity to accommodate peak demand will leave large windows of severe underutilization. Even the best compromise represents unwelcome cost and inefficiency.

With the ability to deploy virtual environments quickly and automatically and redirect capacity as needed, cloud computing offers an ideal solution for testing and development. Cloud vendors makes the transition even more appealing with solutions that allow your clients to experiment with cloud in what is already, by definition, an experimental environment – a low risk introduction to cloud and a first step toward addressing other IT challenges across the business.

The cloud testing services gives the overall business transformation value that helps you reduce the cost associated with IT operations by helping you prioritize your business requirements. A strategic roadmap is required to enjoy the benefits of application virtualization for testing in cloud environment and requires different assessments.

The first assessment is to consolidate the server and conduct the comprehensive virtualization assessment. Getting the requirements of cloud infrastructure is a very difficult task; hence the first assessment provides an opportunity to study the initial requirements of the cloud environments that help to provision automatically, and give the reason to adopt cloud infrastructure and appreciate best practices of cloud. The other assessment is to determine what type of software models can be applied to the available infrastructure to win over the constraints and increase the schedules.

Any business that relies on software development for revenue probably has some benefits to gain from reducing their cycle time and improving their quality. This offering, through its automation and change and configuration management, can reduce errors and speed up deployment.

Businesses with a dedicated development and test organization can benefit from capital and expense reduction afforded by automation and utilization improvements. There is a growing number of organizations that no longer have datacenter space or power to expand their datacenter. By using standardization, consolidation, and virtualization, they can maximize use of their existing assets.

Businesses that are looking to change their business model to be more line-of-business focused may, instead of doing straight allocation, want to do chargeback and have customers pay for only what they use. They may also be able to smooth out resources across different departments so that each department does not have to do account for its own peak demand, and smooth that demand across the peaks and valleys of different departments and different projects.

5.4.4 *Cloud Offering Key Themes*

Nowadays enterprise's datacenter is managed by Operations organizations (likely composed of Infrastructure professionals or System Administrators) teams. Operations departments focus on availability, stability of IT services, and IT cost efficiency. Operations department's goal is to minimize risk for delivering on non-functional requirements by avoiding unnecessary change

and promoting standard infrastructure requirements for applications. Adoption of virtualization technologies has brought a significant change in the enterprise datacenter over the last several years, as it has fundamentally started to change the way IT is delivered and serviced. Enterprise IT Operations managers are tasked with serving two main constituencies:

- **Application teams:** Delivery of production internal and/or external applications according to service level requirements in a cost-effective manner. Application infrastructures are often deployed in silos and provisioned for peak demand, resulting in capacity capabilities far beyond their normal requirements.
- **Development teams:** Development departments are usually driven by user needs for frequent delivery of new features. Development teams often want to test new ideas and/or features quickly in a realistic environment. However, there is often a significant delay between requests to IT for new environments and actual delivery (often can be several weeks). Development teams frequently request and then hesitate to return IT environments back into the centralized pool due to fear of losing access to resources needed for their development cycles.

Due to the history of delivering dedicated environments to support both groups, infrastructure resource utilization metrics are typically below targets. No longer are IT infrastructure managers being pushed solely to reduce cost but also to improve end-user responsiveness. The operations team wants to become more responsive to business needs and reduce application provisioning time by some great percentage (example 90 percent) and increase resource utilization from what are typically very low percentages <20 percent to an exponentially greater reduction(50–60 percent or higher as an example), reducing IT operational costs by some percentage (example 25 percent or more); all while escaping vendor lock-in and regaining control of their applications and infrastructure so they can use it in the most efficient manner to meet the needs of the business.

Virtualization has its benefits, and it does modestly improve resource utilization and delivery times within its hypervisor domain, but nearly all datacenters have multiple hypervisors in use as well as many other computing resources that are not virtualized. Operations teams face significant challenges with manual provisioning and management, VM sprawl, and difficulty scaling when needed.

Key Themes

- **Infrastructure to Applications**
  - The most common use case for private cloud is Infrastructure-as-a-service (IaaS). This can often be as simple as managing VM images to prevent VM sprawl. The vision is to support complex multi-tier application provisioning such that the applications can be fully configured and ready-to-run. In some cases these are then delivered as a Software-as-a-Service (SaaS) model. In between, enterprises create their own Platform-as-a-Service (PaaS) catalogue to enforce consistency in development platforms and middleware. Since companies are using this for their own internal use, they do not need to rewrite their internal applications to match an external PaaS's custom API set, as is required if they were to consider a public PaaS service.
- **Dev/Test to Production**
  - In software development, there is a high level lifecycle for applications. The beginning is development (dev), then test/QA, staging, followed by production.

Enterprises can also differentiate between the types of productions applications between production internal versus external (customer facing) in terms of criticality. In dev/test, the orientation is towards many users with simpler requests (e.g., VMs or infrastructure) with a focus on shorter duration resource usage (reservation, quota management, and reclamation are key to prevent sprawl). Production applications, in contrast, are more concerned about meeting runtime performance SLAs with dynamic scaling and managing more complex, multi-tier applications. The solution should be unique in supporting the full lifecycle in one product offering in the enterprise environment (Figure 5.3).

- **Allocation and Runtime Scaling**
  - Allocation is the process of instantiating the service catalogue item – infrastructure, platforms, or applications. Runtime scaling is the flexing up or down of required resource elements to meet SLA requirements defined by the application owner according to standard corporate standards and business policies. The solution should be unique in providing both IaaS and runtime scaling/workload management in one product offering for enterprise environment.

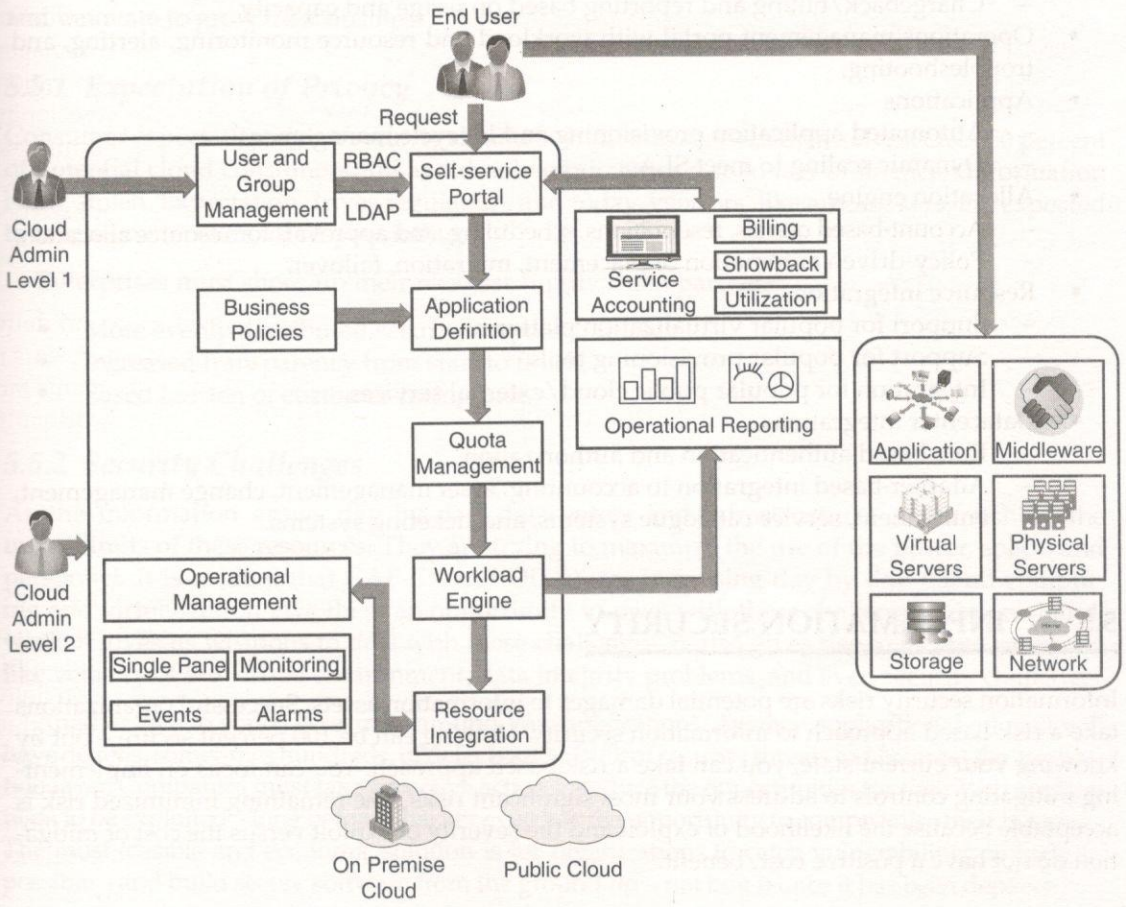


FIGURE 5.3 Cloud orchestration workflow.

Benefits

- **Increase agility and innovation**
  - Enable self-service delivery (minutes).
  - Deliver on SLAs.
  - Simplify process for 'what-if' experimentation.
  - Gain control over public cloud usage.
- **Decrease costs**
  - Increase utilization.
  - Increase operational efficiency (100 servers per admin).
  - Achieve a greener datacenter.
  - Maintain vendor choice.

Offering Key Characteristics

- Service layer
  - Self-service portal for different cloud users: Administrators, Cloud Delegates, and End-Users.
  - Chargeback/billing and reporting based on usage and capacity.
- Operations management portal with workload and resource monitoring, alerting, and troubleshooting.
- Applications
  - Automated application provisioning and lifecycle management.
  - Dynamic scaling to meet SLAs.
- Allocation engine
  - Account-based quotas, reservations, scheduling, and approvals for resource allocation.
  - Policy-driven automation of placement, migration, failover.
- Resource integrations
  - Support for popular virtualization platforms.
  - Support for popular provisioning tools.
  - Integration for popular public cloud/external services.
- Datacenter integrations
  - Role-based authentication and authorization.
  - Adapter-based integration to accounting, asset management, change management, entitlement, service catalogue systems, and ticketing systems.

5.5 INFORMATION SECURITY

Information security risks are potential damages to information assets. Successful organizations take a risk-based approach to information security. Nothing can be 100 percent secure – but by knowing your current state, you can take a risk-based approach. You can focus on implementing mitigating controls to address your most significant risks. The remaining minimized risk is acceptable because the likelihood of exploit and the severity of exploit versus the cost of mitigation do not have a positive cost/benefit.

Risk can be quantified by the expected (average) damage:

- **Value of asset:** What are your valuable information assets?
- **Vulnerabilities:** What vulnerabilities exist in your systems that can be exploited and lead to damage of your assets?
- **Threats:** The level of threats that aim at exploiting vulnerabilities.

Security controls are safeguards or countermeasures to avoid or minimize information security risks:

- **Must be effective:** Mitigate the given risk.
- **Should be adaptive:** Adapt to changing risks.

Three main types of controls:

- **Preventive:** Prevent security incidents (e.g., patching a vulnerability).
- **Detective:** Detect a security incident (e.g., monitoring).
- **Corrective:** Repair damages (e.g., virus removal).

Successful organizations recognize risks, implement the appropriate mitigating controls, and innovate to grow their business.

5.5.1 *Expectation of Privacy*

Consumer expectation is that security should be built into services themselves. Over 50 percent of potential cloud consumers still avoid online purchases due to fear of financial information being stolen. Expectation drives regulation, and today, vendors, like automakers, are expected to take a greater share of responsibility.

Enterprises must shore-up their weakest supply chain partners:

- More evenly distributed security responsibilities.
- Increased transparency from start to finish.
- Eased burden of customer-facing unit.

5.5.2 *Security Challenges*

As the information grows day by day, datacenters and infrastructures are stretching the upper limits of these resources. They are trying to maximize the use of the power, space, and personnel. It is evident that CAP-EX and OP-EX are increasing day by day. Cloud computing and virtualization give them an opportunity to meet with these challenges. On one hand, while it gives us weapons to deal with these challenges, it also gives rise to its own problems like veracity of the virtual environment, data integrity problems, and even security challenge.

Another area to watch is security around Web applications. Average application deployed will have dozens, sometimes hundreds, of defects and a bulk of security threats today target the application layer. Companies must take proactive action to reduce the opportunities for their Web applications to be exploited – long before a hacker even has the opportunity to compromise their business. The most feasible and economic solution is for organizations to catch vulnerabilities as early as possible – and build secure software from the ground up – not bolt it once it has been deployed.

We are moving towards a future where enterprises will adopt the services from cloud service providers externally. The most important point is that the workloads that will use these services will be rendering the low-risk workloads. This will also account for some of the assurances for security, and price of the service will be the deciding factor to whether adopt or not. The workloads that possess medium-to-high risk factors will adopt the private and hybrid cloud. This will also cover the workloads that contain proprietary contents and also those that need more security and depth of defence. Once these services mature and settle, the latter ones will also move towards the external cloud to enjoy the benefits of the external cloud but without compromising the security.

5.5.3 *Security Compliance*

There is a need of policies and procedures for governance and risk factors with respect to cloud security. These services should include procedure to handle change management, incident management, etc. It generates reports for multi-tenant environments. So, we have to bank on large log and audit files to do so. Transparency is an important factor as it is very important for public clouds and it is a black box for the service users.

It is also required to conduct third-party-based checks and audits for the agreements that are breached in the process. Also, third-party-based audits can issue the non-compliance to the subjected violations. This maintains the visibility in the system.

Another method is to have strong Service Level Agreements (SLAs) so that flexibility can be managed for the process based on situations that will enable the traditional outsourcing model and service management to enjoy the benefits of cloud.

5.5.4 *Identity-Based Protection*

Cloud environment requires extra protection levels as it works with diverse set of groups. Therefore, it is essential to have proper authentication for getting access to resources for the environment. It also requires regulated monitoring of users, details regarding the logging to the resources, and check-up for the background verifications. One of the important aspects is to maintain the access that matches with the profile of the work and gauges the risk if something goes wrong due to the improper use of resources. There can be different classes of users like administrators who require the access based on the work they are doing with the cloud environment.

Maintenance of the identity is required to conduct smooth operation in the cloud deployments and authenticate the real users. This is required for both internal and external purposes for the hosted applications. The biggest problem is to make confidential data secure. In order to do this, we have to maintain a secure protocol over the networks, and firewalls should be active to ensure the security of the confidential information; information that is sensitive but not important for the business should be destroyed.

5.5.5 *Data Protection@Cloud*

The relevant terms that dictate the protection of data are how it is stored, how it is accessed, what the compliances are, and what audits are required as per the SLAs. It also relates to regulation of the breaching of the data and its separation on the storage infrastructure. This even includes data that is archived.

This is handled by encryption and managed by encryption keys, and data is protected in the cloud datacenter. Another point that is not taken care of most of the times is the protection of the mobile data. It should be ensured that encryption is done for the mobile data. One of the biggest problems in Internet-based cloud is sending the large amount of data and that is not possible with the Internet-based environments. Therefore, the data should be encrypted and both the cloud service provider and subscriber have the keys to encrypt it.

The movement of the data between the different locations of the organization depends on the cloud environment, support, SLAs, and business activities. There can be violations of the intellectual property law and we should keep this in mind while working with different types of data. It must be ensured that the legal teams should review all the requirements of the cloud environments and how to control the data that is collocated in the large geographic area.

Other important thing that can be classified is the data type and its associated risks for the protection of data. We can have the risks levels and matrix breaches and we can think about security mechanisms based on them. The measures can be different from domain to domain; for instance, public services challenges will be different from financial services data.

5.5.6 *Application Security@Cloud Deployment*

If you think that the protection mechanism in cloud environment is only there at application level, it is wrong. Actually, it is required for the image level. Therefore, cloud vendors should have a clear and sound way to tackle this by meeting the demands of the subscriber for issuing the licenses for the required period of interval, destroying then after use, and making sure that the sensitive unimportant data is also destroyed at the same time.

Following the standards that cloud subscribers demand is one of the important values to the customer for maintaining and supporting the security. All the Web-based requirements should be coded to match the actual requirements, and publishing of the content on the Web should adhere to the policies of the business.

In order to work with the successful protected virtual environments, everybody in the cloud deployment should adhere to an agreed-upon basic security policy. Cloud vendors and subscribers should have the audits check on the intrusion-based polices and check the prevention system put in place to handle this. This becomes more valuable when we work in shared environment because different subscribers on the same the cloud environment should have the agreement for the security and protection policies.

So far, we have talked about the security measures based on software, policies, SLAs, audits, etc., but we should take care of the security in physical terms as well. These security measures can include biometric and security through closed circuit television (CCTV) monitoring. This can restrict unauthenticated entry.

5.6 VIRTUAL DESKTOP INFRASTRUCTURE

Virtual desktop infrastructure provides end-user virtualization solutions. This is designed to help transform distributed IT architectures into virtualized, open-standards-based frameworks leveraging centralized IT services. Virtual desktop infrastructure combines hardware,

software, and services to connect your clients' authorized users to platform-independent, centrally managed applications and full desktop images running as virtual machines running on servers in the datacenter.

The notion behind the virtual desktop infrastructure is to run desktop operating systems and applications inside virtual machines that reside on servers in the datacenter. This is referred to as virtual desktops. Users access virtual desktops and applications from a desktop PC client or thin client using a remote display protocol. The Desktop users get almost all local desktop features as if the applications were loaded on their local systems, the difference being that the applications are centrally hosted and managed.

This solution:

- Consists of optional portal interface, thin-client, and PC with client components or Web browsers with client messaging and security technologies, delivered through a single, consistent framework.
- Delivers a common, standards-based, resilient IT infrastructure that is security-rich and scalable and provides authorized users with single-point, consistent access from a wide choice of client devices.

Project-based services provide IT consultants specializing in virtualization technologies, assistance to assess the organization's desktop and application needs, and subsequently develop a virtual desktop infrastructure solution that best meets these needs.

5.6.1 *Architecture Overview*

Desktop cloud virtualization services provide several advantages to the enterprise. One very important advantage is the reduction of cost. By moving the core function of distributed end-user devices and applications to a centralized infrastructure, the lifecycle of the end-user devices is extended, and the performance requirements are moved to a centralized infrastructure. The administration of a centralized IT infrastructure is more cost efficient than the administration of a distributed one. The implementation of virtualization with virtual desktop infrastructure solutions allows businesses to simplify their IT environment, reducing cost and complexity through consolidation of physical resources and standardization of operating environments.

Virtual desktop infrastructure creates a framework that offers many advantages to the enterprise such as:

- **Cost reduction:** More efficient use of resources can increase utilization.
- **Flexibility:** Common physical infrastructure can support a variety of end-users. New desktop images can be created dynamically without a hardware procurement cycle. Multiple types of guest OSs can be run on the virtual machines so that the physical hardware can support a wide range of end users without costly integration or reconfiguring the systems between user accesses.
- **Security:** The data remains in the datacenter with access control.
- **Availability:** Higher availability as VM can be quickly migrated to a different physical server in the event of a hardware failure.
- **Efficiency:** Service delivery is more efficient when IT processes are optimized for a centralized environment.

5.6.2 *Enterprise Level*

Virtual desktop infrastructure provides a set of proven integration patterns and methods for implementing a client virtualization. The virtual desktop infrastructure team works with various tools and products to help users with an assessment of their environment to develop virtual desktop infrastructure solution requirements and a solution design. The team then develops and deploys this solution into the client environment based on a common architecture that supports the four virtual desktop infrastructure solution models.

Virtual desktop infrastructure solution (shared service, virtual client, workstation blades, streaming) is shaped by the component selection. The base virtual desktop infrastructure architecture is designed to integrate with existing client environments and, as the name implies, provide an access method to a highly scalable virtual infrastructure to the front-end clients. The virtual desktop infrastructure server farm's back-end integrates with existing infrastructure services and legacy applications.

Virtual desktop infrastructure solution introduces a new method of delivering and managing user desktop environments. Virtual desktop infrastructure is a service offering to create a virtual desktop infrastructure. In the IT industry today, several technology vendors provide components that will enable you to build a Virtual Desktop Infrastructure (VDI). VDI is generically used to reference the collection of products and infrastructure components used to form a virtual desktop solution.

Virtual desktop infrastructure solution is designed to reduce the dependency on distributed PC and laptops. By placing critical applications and data in a centralized datacenter with access from a variety of client device options, virtual desktop infrastructure can significantly reduce the cost and complexity of the management and maintenance of desktop images.

The virtual desktop infrastructure end-user desktops run on virtual machines hosted in a centralized IT infrastructure in the client datacenter. The end-users access their individual desktop image or a pool of desktops through a client access device such as PC client, a thin client, or a Web-based client. Applications run on virtual machines on host servers in the datacenter from resource pools rather than on the local machine.

Additional resources can be easily and quickly added to the IT infrastructure as business requirements arise. The virtual desktop infrastructure solution provides increased security as no data leaves the datacenter. Optional secure encapsulation capabilities can allow network connections to be encrypted.

Virtual desktop infrastructure virtual client solution integrates into the organization's datacenter to leverage existing network and infrastructure services. The solution provides access services to the virtual infrastructure hosted in the virtual desktop infrastructure server farm or datacenter (Figure 5.4).

The desktop client devices can be new or existing thin client devices, PCs with an access client, a Web browser, or various combinations depending on the client environment. The desktop client devices can vary from organization to organization, and can be provided as a part of the virtual desktop infrastructure service engagement if required.

The virtual desktop infrastructure access service or 'connection broker' provides device and user authorization, portal integration, session management, host monitoring, application

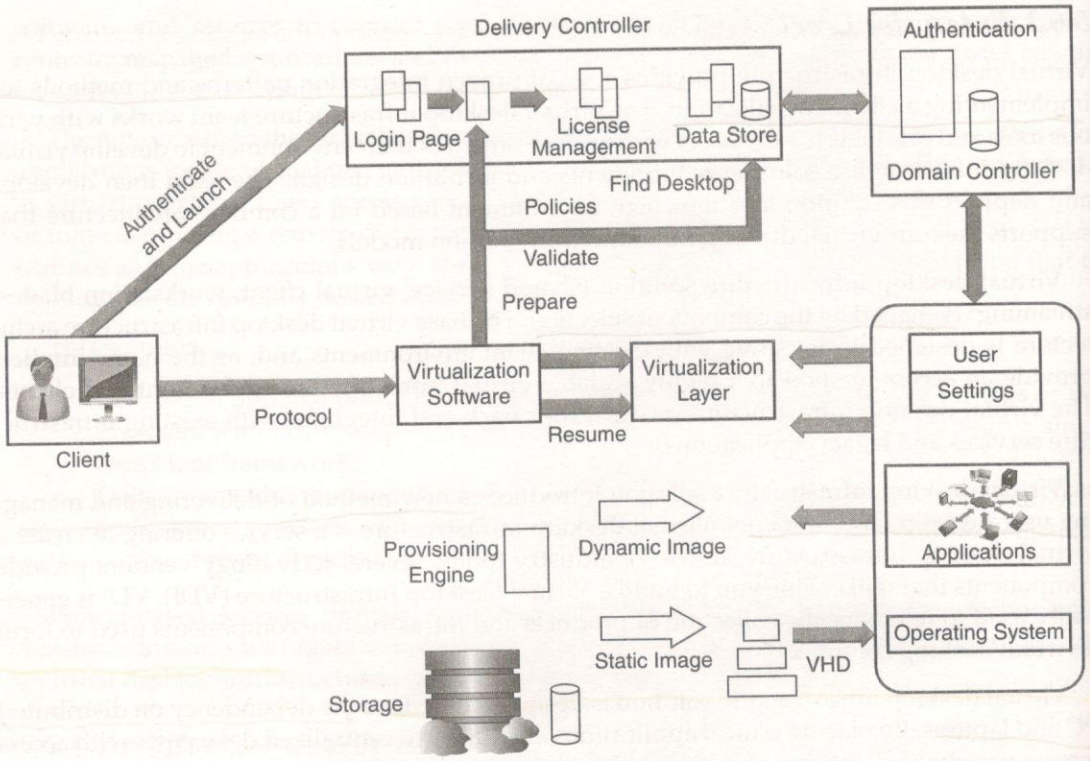


FIGURE 5.4 Virtual desktop infrastructure.

streaming, and consumption-based metering. This also provides load balancing as needed for the front-end security servers, and for the connection servers. The security servers and connection brokers must have high availability, as they are single points of access to the virtual environment.

The virtual desktop infrastructure server farms, or datacenters, consist of a set of physical hardware platforms that facilitates the virtual environment to host shared OS, virtual machines, and dedicated desktop clients.

Infrastructure services include existing services such as activity directory, file, print, management, network, authentication services, and storage. This solution integrates with these existing services rather than introducing redundant features not required by the business. The service will not bundle in components that are available in the business's environment. If additional infrastructure services are required, they can be added to the project-based service for an additional charge. The cloud service project team integrates the virtual desktop infrastructure solution with the existing infrastructure services and desktop client devices as part of the project-based service engagement. Additional offerings may be combined with the cloud service product to meet business requirements as needed.

The virtual desktops images can be configured to have access to the existing infrastructure services in the client datacenter as needed. The following sections briefly discuss the components of the virtual desktop infrastructure at a high level.

5.6.3 Client Access

The users use their remote desktop client device to connect to their virtual desktop. This service is provided mostly by all virtual desktop vendors and supported by a set of products that connects remote clients to the centralized virtual desktops. This process is generically known as connection brokering. The logic controlling which virtual desktop a client should connect is handled by VDI product solution. This makes the process of connecting toVDI Environment simple for the end-user and tightly controlled for the IT administrator.

End-users can use existing PC running operating systems where the user initiates a remote session to a VDI resource using a remote desktop client application or a Web browser.

With some third-party products, you might have local desktop icons configured to access published applications or published desktops.

5.6.4 Desktop Virtualization Services

The core of the virtual desktop infrastructure can be viewed as a central-server-based resource pool with components connecting end-users to applications, networking, and storage resources.

Virtual desktop infrastructure uses vendors to centrally host and deliver a cost-efficient desktop environment from the datacenter, and uses Management server to provide the virtual desktop management.

5.6.5 Desktop Management

Virtualized clients and desktop management proactively manage diverse desktop environments and virtual desktop infrastructure server-based client technology. The end-user retains the features and flexibility of the traditional desktop.

VDI brings together the desirable features of traditional terminal server while retaining important features of distributed computing.

5.6.6 Pool Management for Virtual Desktop Infrastructure

Management server authenticates the user, determines the pool they belong to, and using predetermined policies, provisions a desktop for that end-user, complete with that user's specifications and privileges, and finally deploys it. These pools can be persistent or non-persistent. Non-persistent pools contain multiple hosted virtual desktops, which are initially identical and cloned from the same template. The connection server allocates entitled users to a virtual desktop from the non-persistent pool as requested. This allocation is not retained when the user logs off the desktop and the virtual desktop is placed back into the non-persistent pool for re-allocation to other entitled users. When the user connects to the non-persistent pool on

subsequent occasions the management Connection Server connects the user to any virtual desktop in the non-persistent pool. The persistent pool contains multiple hosted virtual desktops, which are initially identical because they are cloned from the same template. This is typically a many-to-many relationship.

When a group of users is entitled to the persistent pool, every user in the group is entitled to any of the virtual desktops in the pool. The management connection server will allocate users to a virtual desktop as requested. This allocation is retained for subsequent connections.

Individual desktop assignment is a static, one-on-one relationship between a user and a specific virtual desktop. This configuration is good for power users where the desktop is specifically configured for a particular user. This configuration can include specific applications, data access, and resource allocations. Individual desktops give users a high degree of customization.

Maximizing the use of non-persistent pools for all users who do not have a desktop customization requirement is recommended. Typically, the task users could do with a non-persistent desktop.

A virtualized IT environment helps provide security-rich, 'anytime, anywhere' access to applications, information, and resources. Virtual desktop infrastructure is a unique end-user virtualization solution that helps businesses transform their distributed IT architectures into virtualized, open-standards-based frameworks leveraging centralized IT services. It combines hardware, software, and services to connect authorized users to platform independent, centrally managed applications and full client images running in virtual machines.

Desktop cloud project-based service can substantially reduce total cost of ownership by reducing the effort required for desktop PC deployment and management, software distribution,desk-side support and help-desk required to support and maintain desktops.

Virtual desktop infrastructure also offers managed services for businesses that wish to derive benefits of virtual infrastructure access but lack the necessary skills and expertise required for the ongoing management of the virtual infrastructure. Businesses can avoid significant up-front investment and continuing cost for developing and maintaining the necessary skills, knowledge, and experience in systems management and desktop virtualization technologies.

5.7 STORAGE CLOUD

For any type of cloud deployment, whether a private, public, or hybrid cloud, the environments are built using key foundation building blocks such as servers, storage, applications, and infrastructure. Storage and compute resources scale together, and the failure to manage them efficiently results in failure of the cloud services.

Storage management in cloud can help organizations to address their challenges around data and storage management in their clouds – availability of data at all times, storage resource utilization, application performance, longer restore times, higher storage costs, low productivity of storage personnel, increased risk of data loss and downtime.

5.7.1 Value Proposition

Storage cloud reduces the complexity of managing cloud environments by offering a complete portfolio of automated solutions for managing data and storage infrastructure, enabling better efficiency for business resiliency, reducing costs and improving security, while increasing visibility, control, and automation of the cloud storage infrastructure.

Data and storage management within a cloud environment is a critical necessity to provide a reliable, on-demand service experience while at the same time reducing costs and minimizing risks. Streamlining data to target applications plays an important role which means data has to be available at all times and storage provisioned rapidly to the applications built into the cloud for delivering efficient services. Often, storage administrators spend over 50 percent of their time on manual repetitive tasks. They find it difficult to meet stringent rules that are essential to restoring operations quickly after any disruption (database corruption, virus attack, disaster, hardware failure) in a cloud. Failure to ensure data availability at all times can lead to a significant failure of a cloud service.

Also, proper placement of data on different tiers of storage within the cloud, if done efficiently, helps to minimize the overall costs of hardware, software, and administration.

Cloud vendors offers technologies – storage, hardware, and software – as well as key storage services to support subscribers in their journey to leveraging cloud computing. They can assist in planning, designing, building, deploying, and even managing and maintaining storage solutions – whether on their premise or someone else's.

Cloud technology is helping organizations build a smarter business infrastructure with immense flexibility and scalability – one that could result in improving service levels while reducing capital and operational costs. Today, many organizations in various industries including media, banking, etc. that deal with large amounts of data increasingly are adopting cloud technology to address their needs around delivering faster services, protecting data in real time, seamless communication between employees, partners, and suppliers, for business continuity and, of course, the pressure to become more energy efficient – to be a greener organization.

5.7.2 Challenges

Cloud services rely heavily on keeping the data and applications they are managing available at all times, and the ability to restore operations quickly following any type of data disaster (database corruption, virus attack, hardware failure, local/regional disaster) is essential. Storage management is an important function to ensure that data is available, capacity is provisioned rapidly and storage resources are utilized effectively – cloud administrators often find it difficult to meet all the challenges they face concerning storage and data management:

- Data availability and application performance.
- High capital and operating costs, less return on investment.
- Utilization of storage resources.
- Lack of automation – low productivity of storage personnel with specialists doing mundane tasks.

For customers, the drivers for adopting cloud technologies have been cited as:

- Paying for only what they use.
- Cutting costs.
- Monthly payments instead of all up front.
- Having a standardized system.
- Always having the latest software version, since nothing is installed locally.

5.7.3 *Business Drivers*

- Need for standardization and automation of storage services.
- Need to meet service levels consistently – provisioning on-demand computing capacity and storage capacity.
- Need for simplified management of their storage infrastructure – quick provisioning and redeployment of resources, built-in data reduction capabilities.
- Data security and compliance issues.
- Need to lower costs – lack of upfront capital, lower utilization of hardware resources.
- Recovery Point Objectives (RPOs), Recovery Time Objectives (RTOs).

5.7.4 *Benefits*

- Improve service levels by ensuring data availability and application performance and by quick provisioning through automation.
- Reduce capital and operational expenses by leveraging standardization, automation, virtualization.
- Optimize utilization of storage resources and built-in data reduction capabilities to manage more storage with less hardware.
- Reduce hardware, software, and administration costs with policy-based data storage management.
- Manage risk and streamline compliance through real-time data protection.

5.7.5 *Product/Solutions Overview*

Storage management software and services solutions for cloud help ensure that business and IT are fully aligned and supported by integrated service management. They help deliver a workload-optimized approach and offer a choice of implementation options for superior service delivery with agility and speed.

Cloud vendors offer second-generation storage management technology for cloud environments, delivering faster ROI. Cloud storage services include worldwide capability and capacity to provide integrated cloud service offerings to meet your storage management needs.

Cloud vendors reduce the complexity of managing cloud environments by offering a complete portfolio of automated solutions for managing data and storage infrastructure, enabling better efficiency for business resiliency, reducing costs, and improving security, while increasing visibility, control, and automation of the cloud storage infrastructure. There is the need of the broadest, most scalable, and reliable set of storage solutions available to keep cloud services. They should have the complete portfolio for protecting, managing, and virtualizing the environment.

5.7.6 *Product/Solution Description*

Cloud vendors should offer a complete portfolio of software solutions and services for storage management in cloud, designed to help streamlining of storage resources to support cloud services, protection, and management of data, being able to virtualize the entire storage infrastructure, and offer it as a single resource to the cloud.

Cloud-based systems have brought a new, scalable application delivery service model to the market. They can help clients save money and increase flexibility. However, for any type of cloud deployment, streamlining data to target applications in the cloud plays an important role – cloud services rely heavily on data availability and application performance. Additionally, a cloud environment cannot afford longer downtime – after any disruption, it is critical to restore operations quickly. Whether it is an organization managing its own private compute/storage cloud environments or a managed services providers offering cloud-based services (private or public), a key concern is how well the existing resources are optimized in their infrastructure, to improve end-user experience – be able to provide flexibility, speed, reliability, and efficiency. Data and storage management are critical to improve on-demand service experience, reduce costs, and reduce risks in the cloud.

Often, the absence of sophisticated storage management systems in a cloud results in lack of visibility into the storage utilization and provisioning, costs, and associated risks. Cloud administrators have difficulty in understanding how much capacity is available, where it is, which applications are accessing it, how secure they are, and whether they are able to meet stringent RTOs. Gaps that may exist between the unpredictable demand for data availability and the ability of business to support the same in an efficient way, is resulting in unmet service levels, additional downtime, new hardware and operational costs, and lower customer satisfaction.

5.8 SUMMARY

Today, we are becoming more interconnected, instrumented, and intelligent – the world is becoming smaller. As a result, more data is being generated within the operations of all organizations, and they are struggling with managing the complexity in their storage environments. The costs of backup and recovery, archiving, expiration, and storage resource management are exploding. It is not just about increasing capacity but managing data efficiently, reducing data, ensuring adequate protection of data, and quick restore for better business performance. We need better implementation of virtual desktop infrastructure, test cloud environments, and analytics to handle cloud offerings.

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 of 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.

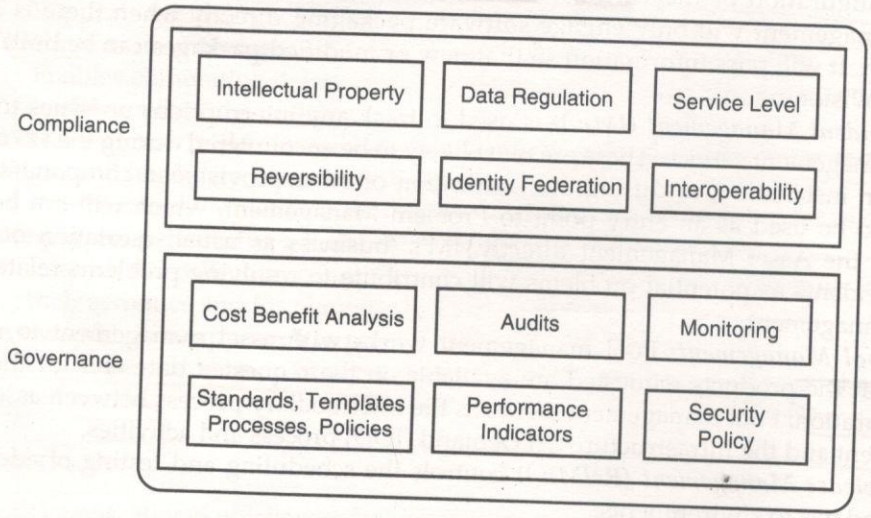


FIGURE 6.1 Compliance and Governance.

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.

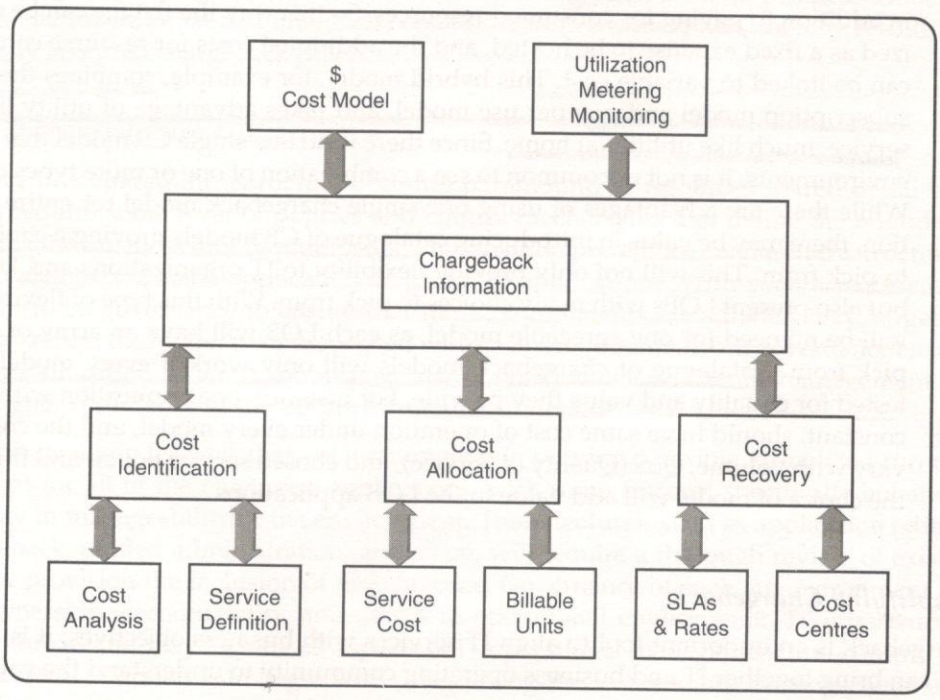


FIGURE 6.2 Chargeback model.

- **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 Model/Market-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.

Resource Based
Activity Based
flat fee
External pricing model
Resource or Usage Based
Product or Service Based

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 a 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

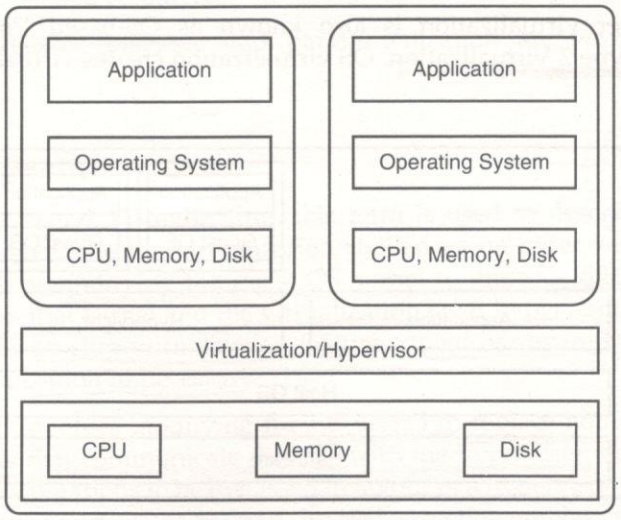


FIGURE 7.1 Server virtualization.

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

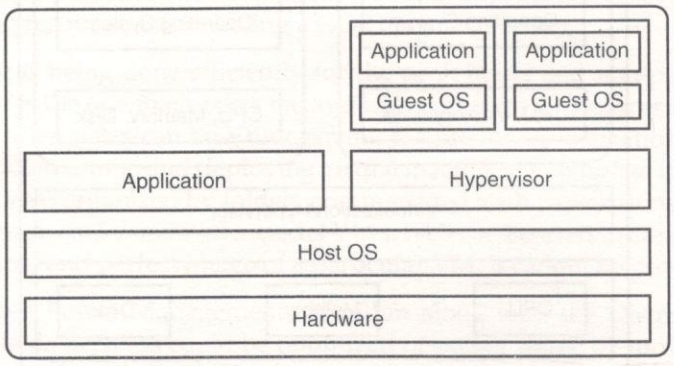


FIGURE 7.2 OS virtualization.

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

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).

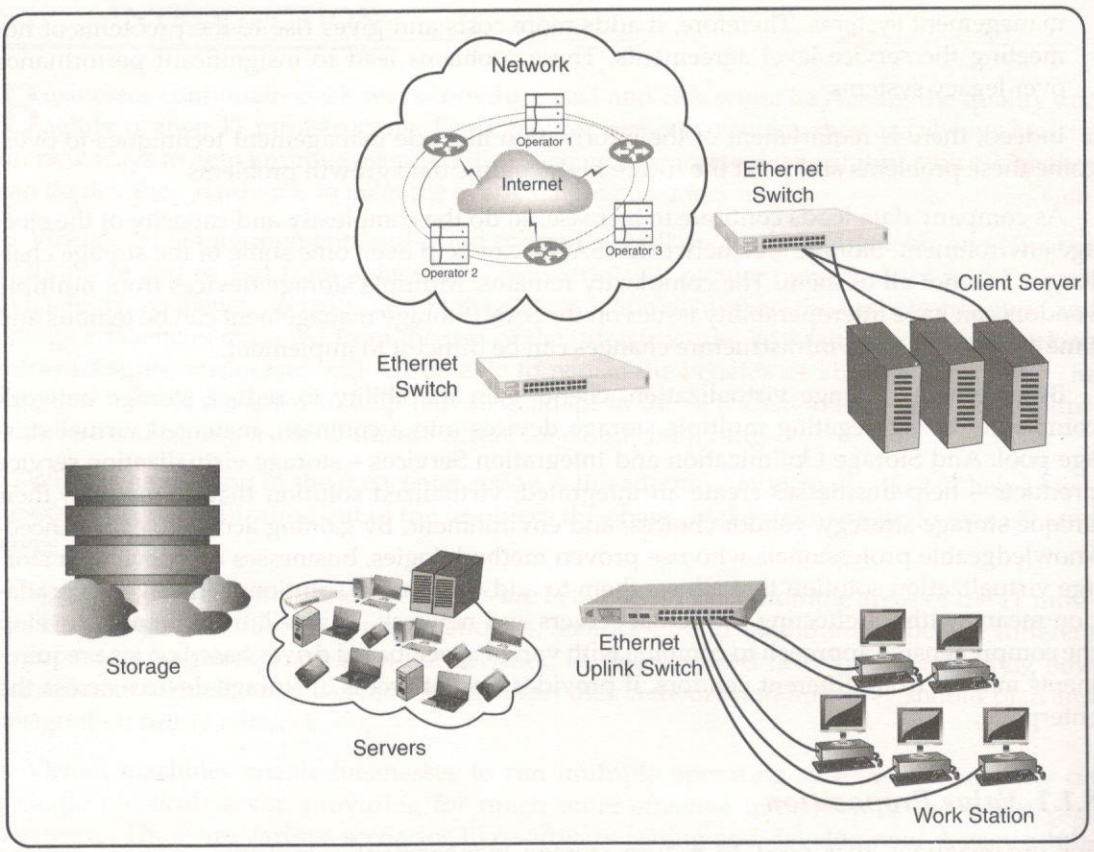


FIGURE 8.1 Storage cloud.

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 storage network, 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.

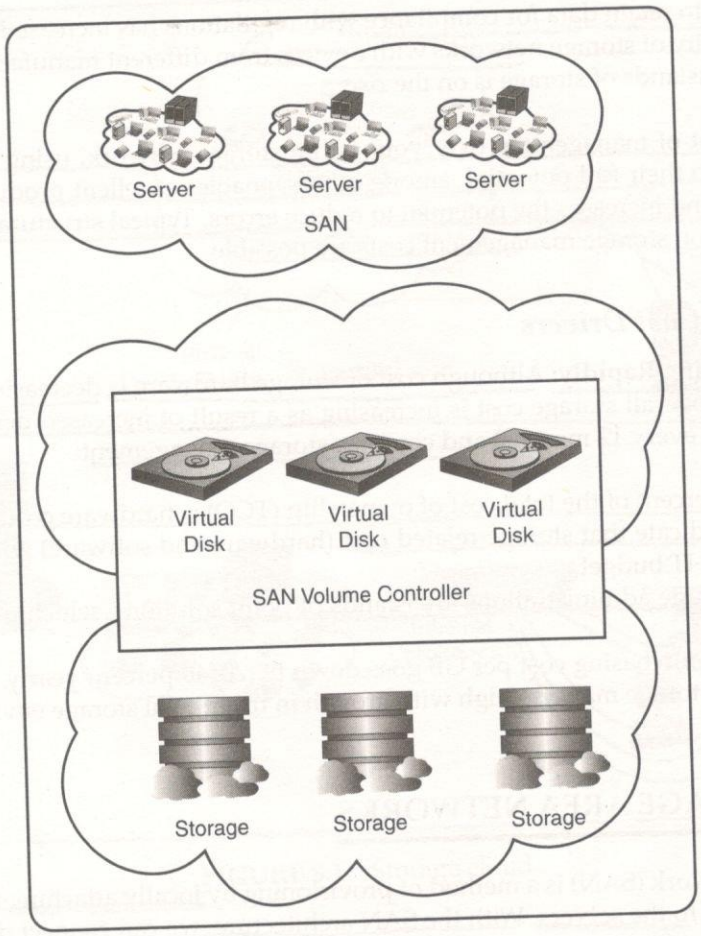


FIGURE 8.2 SAN.

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 servers have 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.

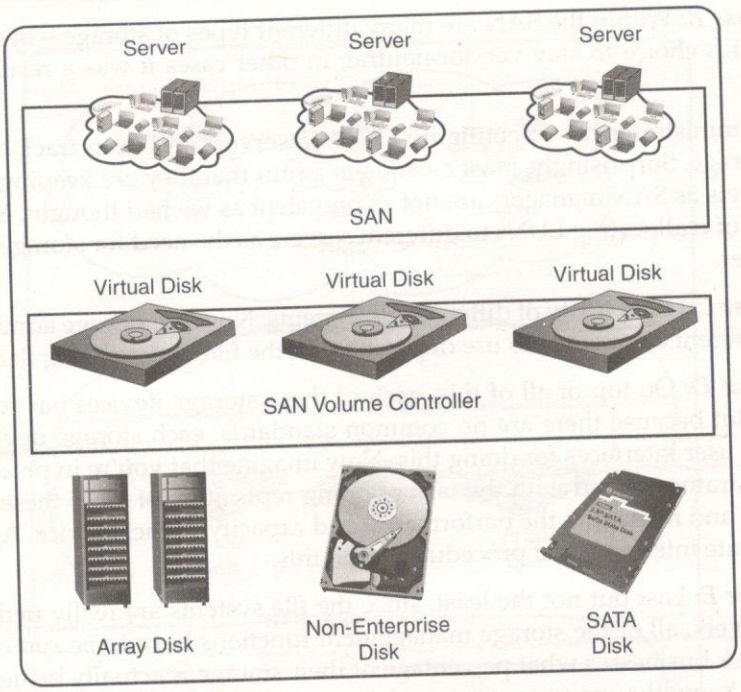


FIGURE 8.3 Virtual storage infrastructure.

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.

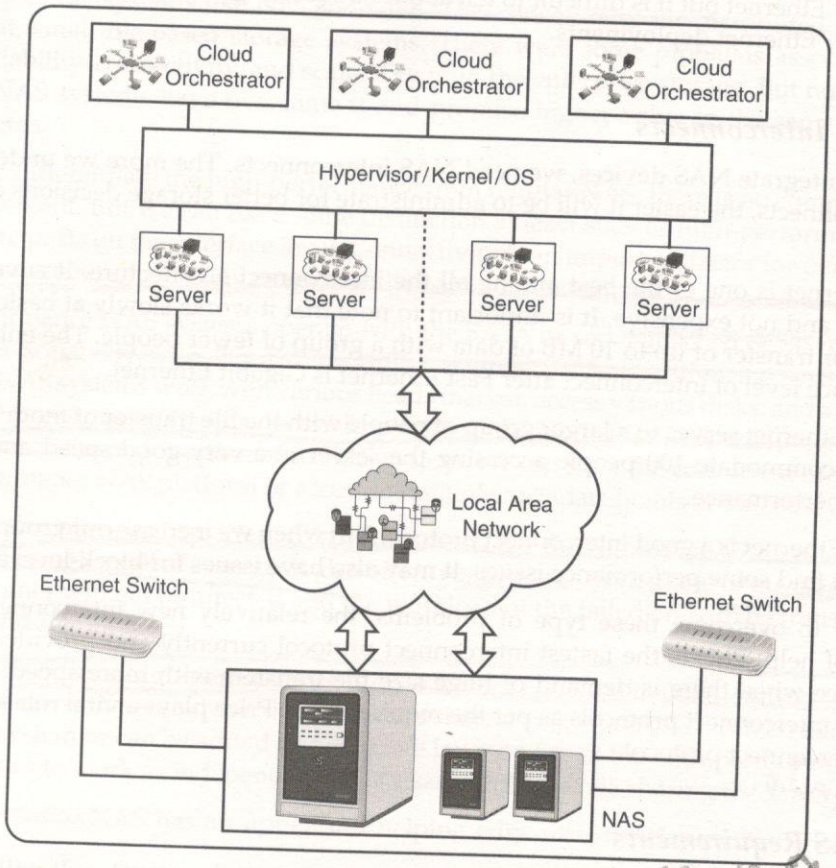


FIGURE 8.4 NAS.

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).

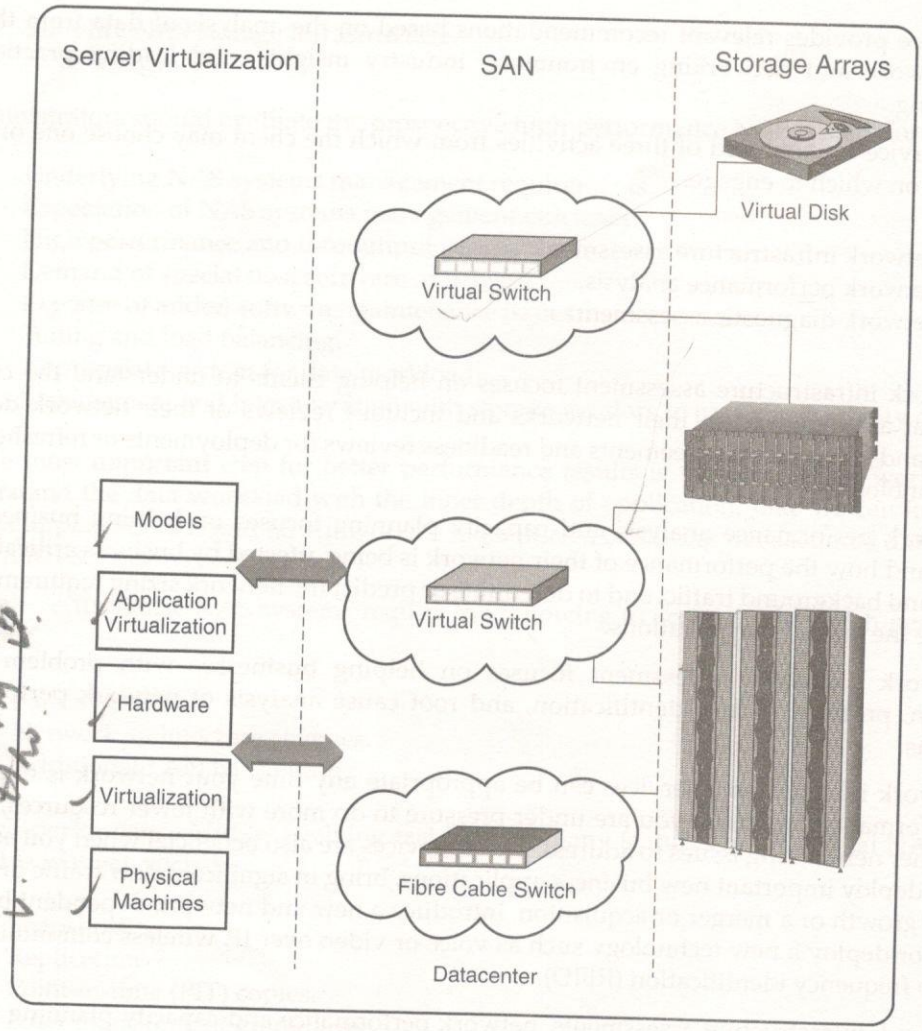


FIGURE 8.5 Server virtualization.

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 virtual zation.
- **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

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).

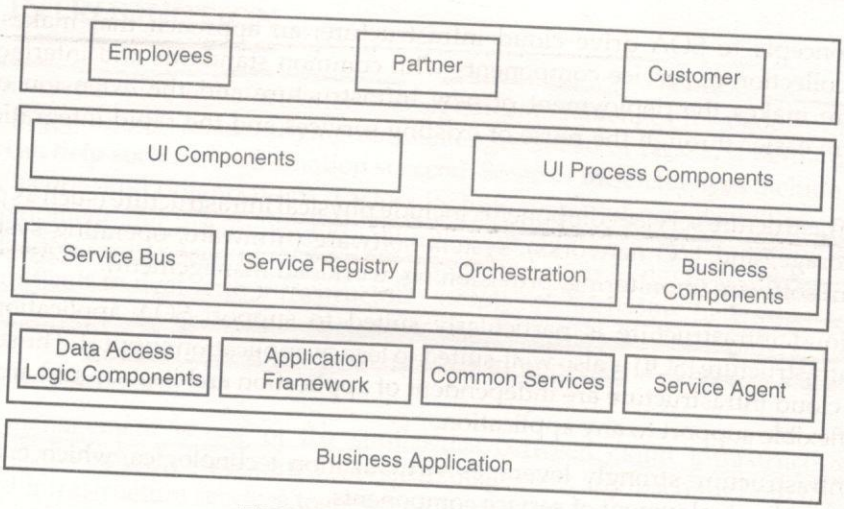


FIGURE 9.1 SOA Model.

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.

SOA →

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.

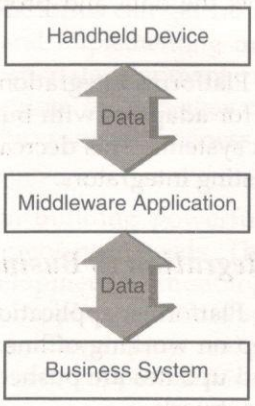


FIGURE 10.1 Mobility application model.

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).

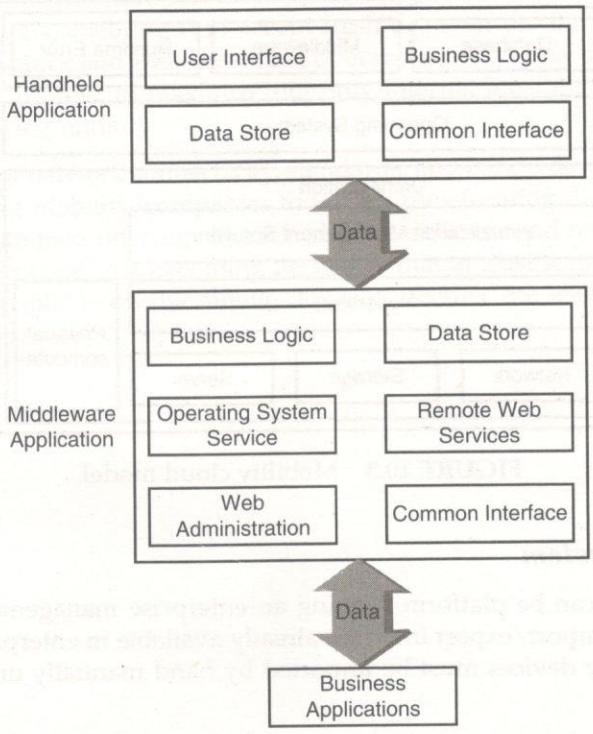


FIGURE 10.2 Mobility application architecture.

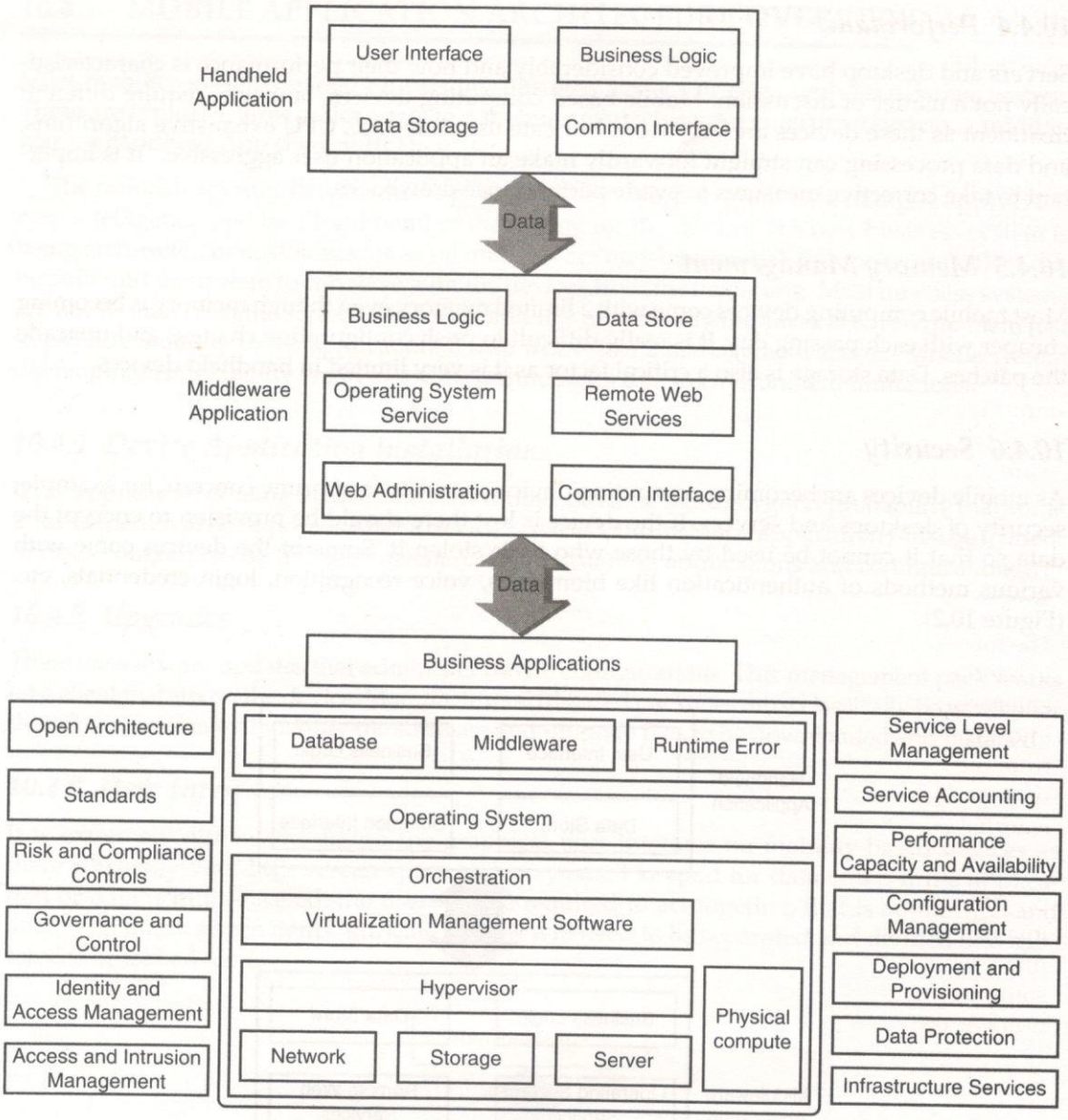


FIGURE 10.3 Mobility cloud model.

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.

APPENDIX A

CLOUD PERFORMANCE MONITORING COMMANDS

This appendix discusses several cloud infrastructure performance data collection and performance monitoring commands.

A.1 `vmstat` COMMAND

`vmstat` command: Account virtual memory information

Syntax

`vmstat [-n]` *`[delay [count]]`*

`vmstat[-V]`

Description

The important factors for virtual memory are paging, swap, memory, traps, blocks, IO and CPU. `vmstat` statistics gives the vital information regarding paging, swap, memory, traps, blocks, IO and CPU. It gives information about the system from the time of the reboot as an average. We can add the sampling delay for the `vmstat` as an option.

We will discuss the different parts of the `vmstat` command.

Process Part

- `r` field: Total number of runnable process.
- `b` field: Total number of blocked process.

Memory Part

- `Swpd` field: Used swap space.
- `Free` field: Available free RAM.
- `Buff` field: RAM used for buffers.
- `Cache` field: RAM used for file system cache.

Swap Part

- `Si` field: Amount of memory swapped from disk per second.
- `So` field: Amount of memory swapped to disk per second.

IO Part

- `Bi` field: Blocks received from disk.
- `Bo` field: Blocks sent to disk.

System Part

- `In` field: Number of interrupts per second.
- `Cs` field: Number of context switches per second.