Unit 1

Chapter 1

Unit Structure

1.0 Objective

This chapter would make your under the concept of following concepts

- What is a cloud computing?
- What are characteristics and benefits of cloud computing?
- It's Challenges.
- Historical development of technologies toward the growth of cloud computing
- Types of Cloud Computing Models.
- Different types of Services in the Cloud Computing.
- Application development and Infrastructure and system development technologies about the Cloud Computing.
- Overview of different sets of Cloud Service Providers.

1.1 Introduction

Historically, computing power was a scarce, costly tool. Today, with the emergence of cloud computing, it is plentiful and inexpensive, causing a profound paradigm shift — a transition from scarcity computing to abundance computing. This computing revolution accelerates the commoditization of products, services and business models and disrupts current information and communications technology (ICT) Industry .It supplied the services in the same way to water, electricity, gas, telephony and other appliances. Cloud Computing offers on-demand computing, storage, software and other IT services with usage-based metered payment. Cloud Computing helps re-invent and transform technological partnerships to improve marketing, simplify and increase security and increasing stakeholder interest and consumer experience while reducing costs. With cloud computing, you don't have to over-provision resources to manage potential peak levels of business operation. Then, you have the resources you really required. You can scale these resources to expand and shrink capability instantly as the business needs evolve. This chapter offers a brief summary of the trend of cloud computing by describing its vision, addressing its key features, and analyzing technical advances that made it possible. The chapter also introduces some key cloud computing technologies and some insights into cloud computing environments.

1.2 Cloud computing at a glance

The notion of computing in the "cloud" goes back to the beginnings of utility computing, a term suggested publicly in 1961 by computer scientist John McCarthy:

"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility… The computer utility could become the basis of a new and important industry."

The chief scientist of the Advanced Research Projects Agency Network (ARPANET), Leonard Kleinrock, said in 1969:

"as of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country."

This vision of the computing utility takes form with cloud computing industry in the 21st century. The delivery of computing services is easily available on demand just like other utilities services such as water, electricity, telephone and gas in today's society are available. Likewise, users (consumers) only have to pay service providers if they have access to computing resources. Instead of maintaining their own computing systems or data centers, customer can lease access from cloud service providers to applications and storage. The advantage of using cloud computing services is that organizations can avoid the upfront cost and difficulty of running and managing their own IT infrastructure and pay for when they use it. Cloud providers can benefit from large economies of scale by offering the same services to a wide variety of customers.

In the case, consumers can access the services according to their requirement with the knowing where all their services are hosted. These model can called as utility computing as cloud computing. As cloud computing called as utility computing because users can access the

infrastructure as a “cloud” as application as services from anywhere part in the world. Hence Cloud computing can be defined as a new dynamic provisioning model of computing services that improves the use of physical resources and data centers is growing uses virtualization and convergence to support multiple different systems that operate on server platforms simultaneously. The output achieved with different placement schemes of virtual machines will differ a lot. .

By observing advancement in several technologies , we can track of cloud computing that is (virtualization, multi-core chips), especially in hardware; Internet (Web services, service-oriented architectures, Web 2.0), Distributed computing (clusters, grids), and autonomous Computing, automation of the data center). The convergence of Figure 1.1 reveals the areas of technology that have evolved and led to the advent Cloud computing. Any of these technologies were considered speculation at an early stage of development; however, they received considerable attention later Academia and big business companies have been prohibited. Therefore, a Process of specification and standardization followed which resulted in maturity and wide adoption. . The rise of cloud computing is closely associated with the maturity of these technologies.

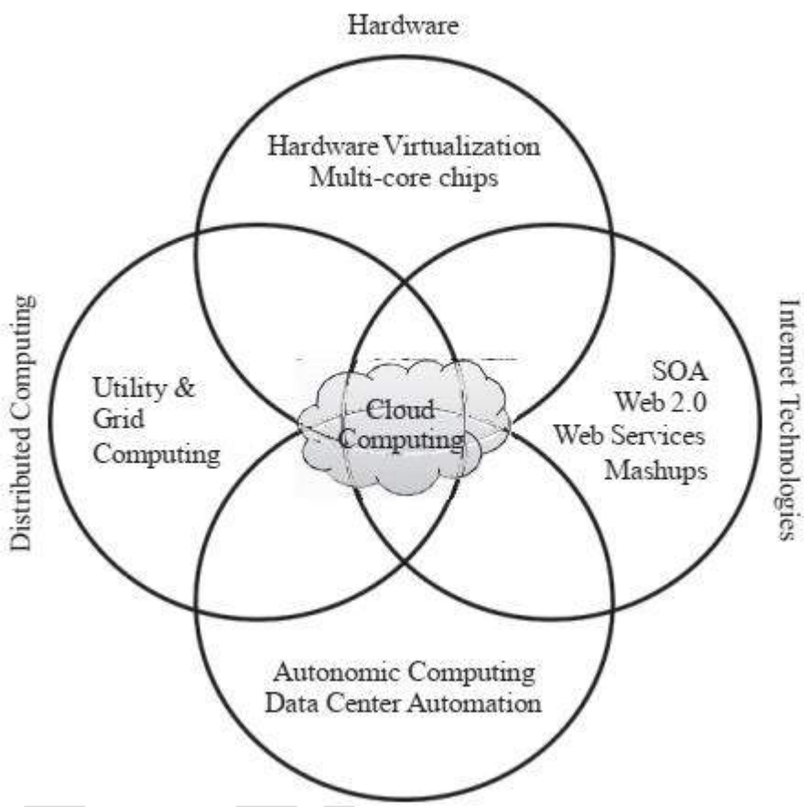


FIGURE 1.1. Convergence of various advances leading to the advent of cloud computing

1.2.1 The vision of cloud computing

The virtual provision of cloud computing is hardware, runtime environment and resources for a user by paying money. As of these items can be used as long as the User, no upfront commitment requirement. The whole computer device collection is turned into a Utilities set that can be supplied and composed in hours rather than days together, to deploy devices without Costs for maintenance. A cloud computer's long-term vision is that IT services are traded without technology and as utilities on an open market as barriers to the rules.

We can hope in the near future that it can be identified the solution that clearly satisfies our needs entering our application on a global digital market services for cloud computing. This market will make it possible to automate the process of discovery and integration with its existing software systems. A digital cloud trading platform is available services will also enable service providers to boost their revenue. A cloud service may also be a competitor's customer service to meet its consumer commitments.

Company and personal data is accessible in structured formats everywhere, which helps us to access and communicate easily on an even larger level. Cloud computing's security and stability will continue to improve, making it even safer with a wide variety of techniques. Instead of concentrating on what services and applications they allow, we do not consider "cloud" to be the

most relevant technology. The combination of the wearable and the bringing your own device (BYOD) with cloud technology with the Internet of Things (IOT) would become a common necessity in person and working life such that cloud technology is overlooked as an enabler.

Figure 1.2. Cloud computing vision.
(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

1.2.2 Defining a cloud

The fairly recent motto in the IT industry "cloud computing," which came into being after many decades of innovation in virtualization, utility computing, distributed computing, networking and software services. A cloud establishes an IT environment invented to provide measured and scalable resources remotely. It has evolved as a modern model for information exchange and internet services. This provides more secure, flexible and scalable services for consumers. It is used as a service-oriented architecture that reduces end-user overhead information.

Figure 1.3 illustrates the variety of terms used in current cloud computing definitions.

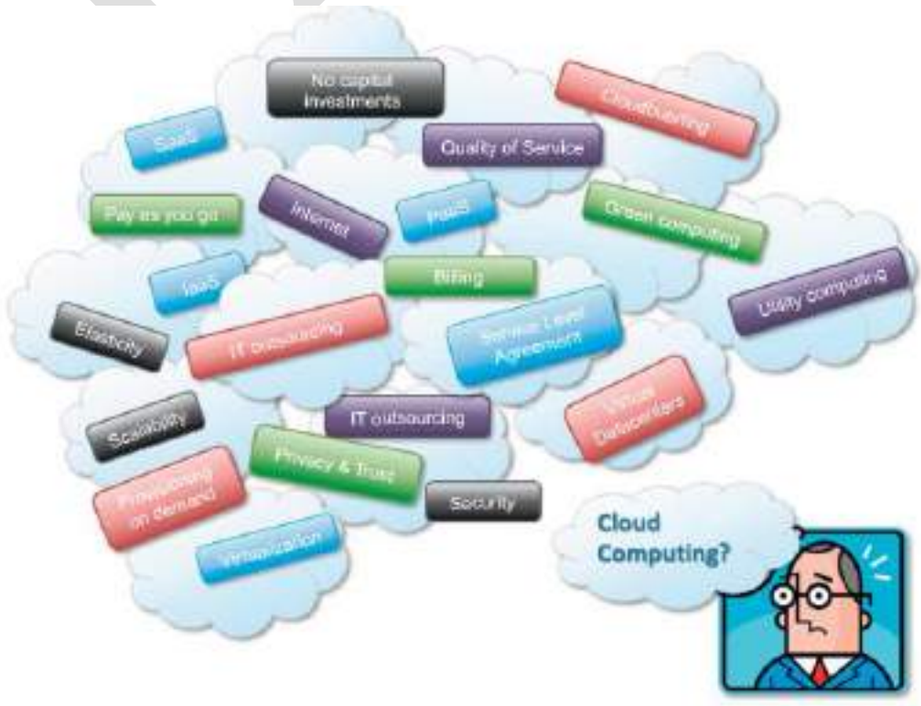


FIGURE 1.3 Cloud computing technologies, concepts, and ideas.

(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

Internet plays a significant role in cloud computing for representing a transportation medium of cloud services which can deliver and accessible to cloud consumer. According to the definition given by Armbrust

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and system software in the datacenters that provide those services

Above definition indicated about the cloud computing which touching upon entire stack from underlying hardware to high level software as service. It introduced with the concept of *everything as service* called as *Xaas* where different part of the system like IT Infrastructure , development platform for an application ,storage ,databases and so on can be delivered as services to the cloud consumers and consumers has to paid for the services what they want. This new paradigms of the technologies not only for the development of the software but also how the user can deploy the application ,make the application accessible and design of IT infrastructure and how this companies allocate the costs for IT needs. This approach encourage the cloud computing form global point of views that one single user can upload the documents in the cloud and on the others side Company owner want to deploy the entire infrastructure in the public cloud. According to the definition proposed by the U.S. National Institute of Standards and Technology (NIST):

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Another approach of cloud computing is "utility computing" where could computing mainly focus on delivering services based upon the pricing model it called as "*pay-per-use*" strategy. Cloud computing make all the resources online mode such as storage, you can lease virtual hardware or you can use the resource for the application development and users has to pay according to their usage their will no or minimal amount of upfront cost. All this above operations are performed and user have to pay the bill by simply entering the credit card details and accesses this services through the web browsers. According to George Reese

He have defined three criteria on whether a particular service is a cloud service:

- The service is accessible via a web browser (nonproprietary) or web services API.
- Zero capital expenditure is necessary to get started.
- You pay only for what you use as you use it.

Many cloud service providers provides the cloud services freely to the users but some enterprise class services can be provided by the cloud service providers based upon specific pricing schemes where users have to subscribe with the service provider on which a service level agreement (SLA) is defined based on the quality parameters between the cloud service providers and user and cloud service providers has to delivered the services according the service level agreement (SLA)

RajKumar Buyya defined cloud computing based on the nature of utility computing

A cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers.

1.2.3 A closer look

Cloud computing is useful in governments, enterprises, public and private institutions and research organizations which make more effective and demand-driven computing services systems. There seem to be a number of specific examples demonstrating emerging applications of cloud computing in both established companies and startups. Such cases are intended to illustrate the value proposition of viable cloud computing solutions and the benefits businesses have gained from these services.

NewYork Times : One of the most widely known examples of cloud computing commitment comes from New York Times . The New York Times has collected a large number of high-resolution scanned images of historical newspapers, ranging from 1851-1922. They want to process this set of images into separate articles in PDF format. Using 100 EC2 instances, they can complete the processing within 24 hours at a total cost of $ 890 (EC2 calculation time is $ 240, S3 data transfer and storage use is $ 650, storage and transfer of 4.0TB source image and 1.5TB Output

PDF). Derek Gottfrid pointed out: "Actually, it worked so well that we ran it twice, because after the completion we found an error in the PDF."

The New York Times had the option to utilize 100 servers for 24 hours at the low standard cost of ten cent an hour for every server. In the event that the New York times had bought even a solitary server for this errand, the probable expense would have surpassed the $890 for simply the hardware, and they likewise need to think about the expense of administration, power and cooling Likewise, the handling would have assumed control more than a quarter of a year with one server. On the off chance that the New York Times had bought four servers, as Derek Gottfrid had considered, it would have still taken almost a month of calculation time. The quick turnaround time (sufficiently quick to run the activity twice) and endlessly lower cost emphatically represents the prevalent estimation of cloud services.

Washington Post : In a related but more latest event, the Washington Post were able to transform 17,481 pages of scanned document images into a searchable database in just a day using Amazon EC2. On March 19th at 10am, Hillary Clinton's official White House schedule from 1993-2001 was published to the public as a large array of scanned photographs (in PDF format, but non-searchable). Washington Post programmer Peter Harkins utilized 200 Amazon EC2 instances to conduct OCR (Optical Character Recognition) on the scanned files to create searchable text – " I used 1,407 hours of virtual machine time with a total cost of $144.62. We find it a positive proof of concept.

DISA : Federal Computer Week mentioned that the Defense Information Systems Agency (DISA) as compared the cost of the usage of Amazon EC2 versus internally maintained servers : "In a latest take a look at, the Defense Information Systems Agency in comparison the price of growing a simple application known as the Tech Early Bird on $30,000 well worth of in-house servers and software program with the costs of growing the equal application using the Amazon Elastic Compute Cloud from Amazon Web Services. Amazon charged 10 cents an hour for the provider, and DISA paid a total of $5 to expand a software that matched the overall performance of the in-house application.

SmugMug : SmugMug, an image posting and hosting web site like Flickr , stores a substantial level of its photo information in Amazon's S3 cloud storage service . In 2006, they ended up saving "$500,000 in prepared disk drive expenses in 2006 and reduce its disk storage space array costs in half" through the use of Amazon S3. Based on the CEO of SmugMug, they might "easily save a lot more than $1 million" in the next year through the use of S3. The CEO known that their present growth rate during the article necessitates about $80,000 worth of new hardware, and the regular costs boost even more considerably after putting "power, cooling, the info center space, along with the manpower had a need to manage them." On the other hand, Amazon S3 costs around $23,000 per month for equivalent storage which is all-inclusive (power, maintenance, cooling, etc. are figured into the expense of the storage.

Eli Lily : Eli Lily, among the largest pharmaceutical companies, is needs to utilize Amazon's storage and compute clouds to supply on-demand high-performance processing for research reasons . John Foley highlights, "it accustomed to acquire Eli Lilly seven and a half weeks to deploy a server internally" whereas Amazon can provision a virtual server in 3 minutes. Furthermore "a 64-node Linux cluster could be online in 5 minutes (compared against 90 days internally)." Amazon's cloud providers not only deliver on-demand scaling and usage-based billing, they enable Eli Lily to respond with considerably amplified agility in addition, eliminating time-consuming products deployment and acquisition functions.

Best Buy's Giftag: Best Buy's Giftag is a new online wish-list service hosted by Google's App Engine. In a video interview, the developers suggested that they were beginning to build a platform with a different technology and moved to Google App Engine with its superior speed of development and scaling advantages. As one developer eloquently stated it, "a lot of the work that none of us even needs to do is [already] completed for us." The developers also lauded App Engine 's design to allow effortless scaling; App Engine-based web apps inherit Google's best-in - class technologies and expertise in running large-scale websites. By the end of the day, App Engine helps developers to focus on building site-specific separated features: "Not worried with the operational aspects of an application going away always frees you to create excellent code or evaluate your code better.

TC3 : TC3 (Total Claims Capture & Control) is a healthcare services company imparting claims management solution. TC3 now makes use of Amazon's cloud services to allow on-demand scaling of resource and lower infrastructure costs . TC3's CTO notes, "we're making use of Amazon S3, EC2, and SQS to permit our claim processing capacity to growth and reduce as required to satisfy our service level agreements (SLAs). There are times we require massive quantities of computing resource that a long way exceed our machine capacities and when these

conditions took place inside the past our natural response became to name our hardware vendor for a quote. Now, by using the usage of AWS products, we can dramatically reduce our processing time from weeks or months right down to days or hours and pay much less than shopping, housing and maintaining the servers ourselves" Another particular feature of TC3 's activities is that, because they provide US health-related services, they are obligated to abide with the HIPPA (Health Insurance Portability and Accountability Act). Regulatory compliance is one of the main obstacles facing corporate adoption of cloud infrastructure – the fact that TC3 is capable of complying with HIPPA on Amazon's platform is significant.

How all of the computing made possible? in same IT services on demand like computing power , storage and providing an runtime environments for development of an applications on pay-as-you go basis .cloud computing not only provides an opportunity for easily accessing of IT services as per demand , but also provides newly ideas regarding IT Services and resources as am utilities .Figure 1.4 provides a bird's-eye view of cloud computing

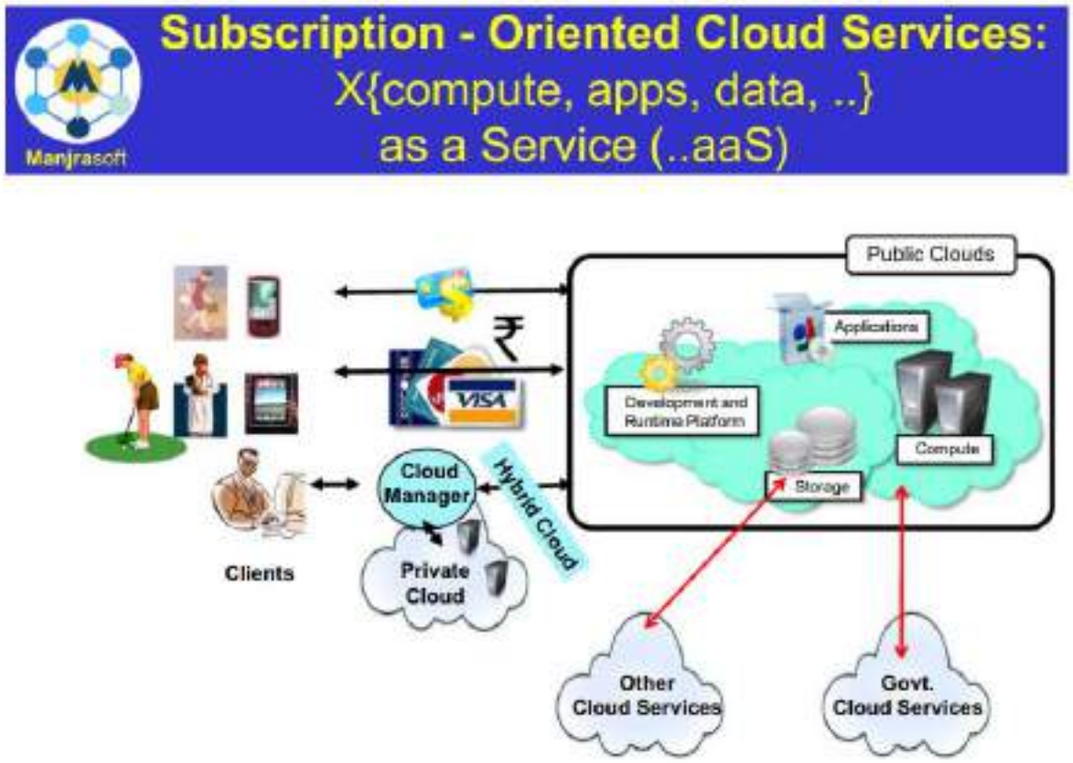


FIGURE 1.4 A bird's-eye view of cloud computing
(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

There are three deployment models for accessing the services of cloud computing environment are public, private and hybrids clouds (see Figure 1.5). The *public cloud* is one of the most common deployment models in which computing services is offered by third-party vendors that the consumer are able to access and purchase the resource from the public cloud via the public internet. These can be free or on-demand, meaning that consumers pay for their CPU cycles, storage or bandwidth per use. Public clouds will save companies from the expensive procurement, management and on-site maintenance of hardware and application infrastructure — all management and maintenance of the system is held to responsibility of the cloud service provider. *Public clouds* can also be deployed faster than on-site infrastructures with a platform almost constantly scalable. Although security issues have been posed by public cloud implementations, the public cloud could be as secure as the most efficiently operated private cloud deployment when it is implemented correctly. A *private cloud* is an essentially one organization's cloud service. In using a *private cloud*, the advantages of cloud computing are experienced without sharing resources with other organizations. There can be a private cloud within an organization, or be controlled from a third party remotely, and accessed via the Internet (but it is not shared with others, unlike a public cloud). Private cloud incorporates several of the advantages of cloud computing — including elasticity, scalability and easy service delivery — with the on-site control, security, and resource customization .Many companies select private cloud over public cloud (cloud computing services delivered through multi-customer infrastructure) because private cloud is a simpler (or the only way) way to satisfy their regulatory compliance requirements. Others prefer private cloud because their workloads deal with confidential information, intellectual property, and personally identifiable information (PII), medical records, financial data and other sensitive data. *Hybrid cloud* is an infrastructure that contains links between a user's cloud (typically referred to as "*private cloud*") and a third-party cloud (typically referred to as "*public cloud*").

Whilst the private and public areas of the hybrid cloud are linked, they remain unique. This allows a hybrid cloud to simultaneously offer the advantages of several implementation models. The sophistication of hybrid clouds is very different. Some hybrid clouds, for example, only connect the on-site to public clouds. The operations and application teams are responsible for all the difficulties inherent in the two different infrastructures.

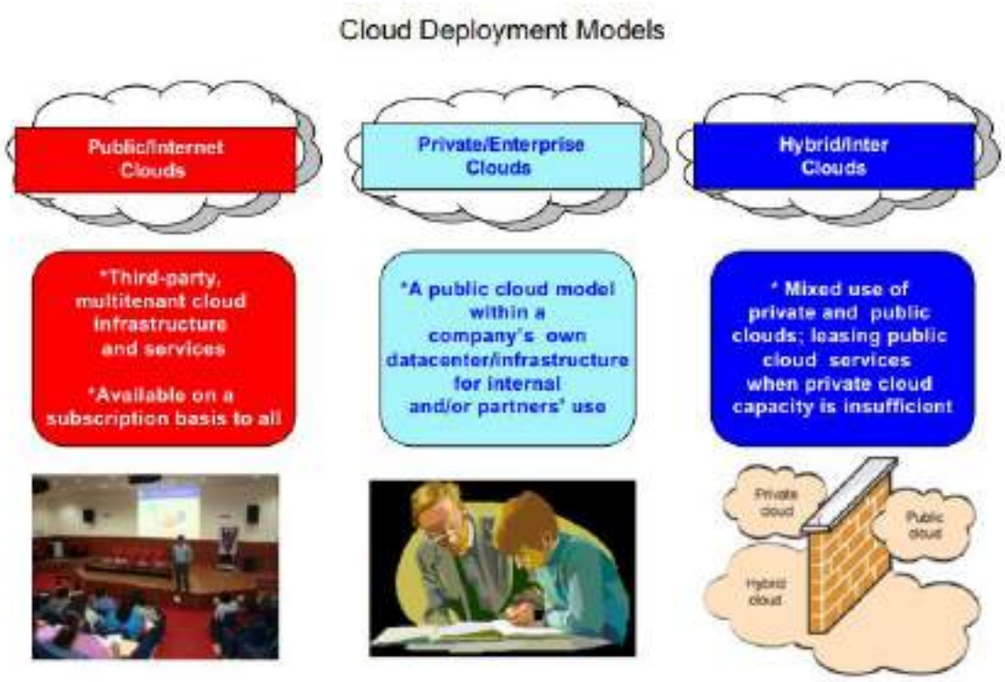


FIGURE 1.5 Major deployment models for cloud computing.
Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

1.2.4 The cloud computing reference model

A model that characterizes and standardizes the functions of a cloud computing environment, is the cloud reference model. This is a basic benchmark for cloud computing development. The growing popularity of cloud computing has expanded the definitions of different cloud computing architectures. The cloud environment has a wide range of vendors and multiple offer definitions which make the evaluation of their services very hard. The way the cloud functions and interacts with other technology can be a little confusing with such complexity in its implementation.

A standard cloud reference model for architects, software engineers, security experts and businesses is required to achieve the potential of cloud computing. This cloud landscape is controlled by the Cloud Reference Model. Figure 1.6 displays various cloud providers and their innovations in the cloud services models available on the market.

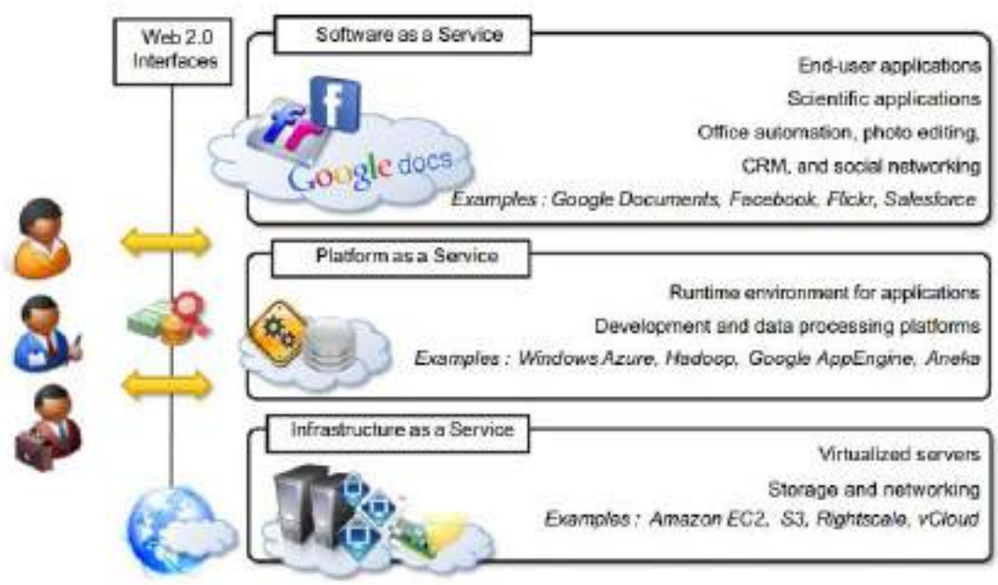


FIGURE 1.6 The Cloud Computing Reference Model.

7. Dealing with Multi-Cloud Environments

Today not even a single cloud is operating with full businesses. According to the RightScale report revelation, almost 84 percent of enterprises adopt a multi-cloud approach and 58 percent have their hybrid cloud approaches mixed with the public and private clouds. In addition, five different public and private clouds are used by organizations.

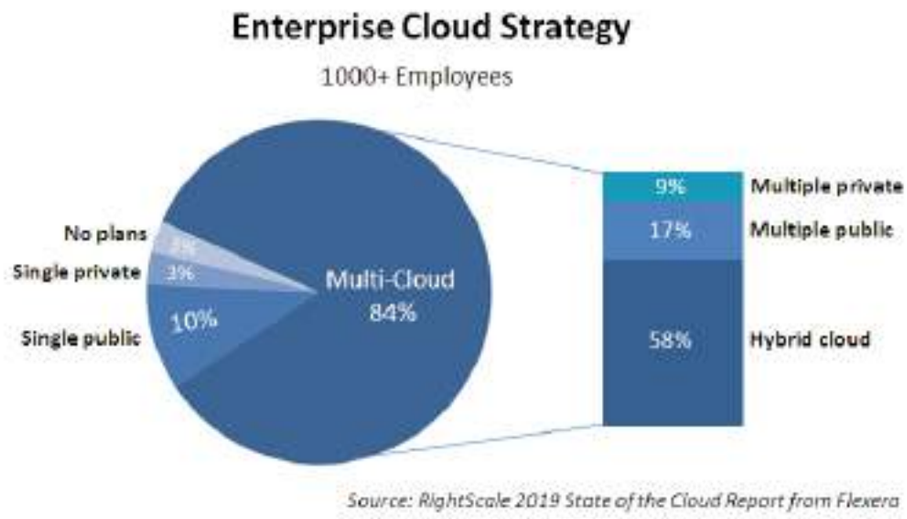


FIGURE 1. 8 RightScale 2019 report revelation

The teams of the IT infrastructure have more difficulty with a long-term prediction about the future of cloud computing technology. Professionals have also suggested top strategies to address this problem, such as rethinking processes, training personnel, tools, active vendor relations management, and the studies.

8. Cloud Migration

While it is very simple to release a new app in the cloud, transferring an existing app to a cloud computing environment is harder. 62% said their cloud migration projects are harder than they expected, according to the report. In addition, 64% of migration projects took longer than expected and 55% surpassed their budgets. In particular, organizations that migrate their applications to the cloud reported migration downtime (37%), data before cutbacks synchronization issues (40%), migration tooling problems that work well (40%), slow migration of data (44%), security configuration issues (40%), and time-consuming troubleshooting (47%). And to solve these problems, close to 42% of the IT experts said that they wanted to see their budget increases and that around 45% of them wanted to work at an in-house professional, 50% wanted to set the project longer, 56% wanted more pre-migration tests.

9. Vendor lock-in

The problem with vendor lock-in cloud computing includes clients being reliant (i.e. locked in) on the implementation of a single Cloud provider and not switching to another vendor without any significant costs, regulatory restrictions or technological incompatibilities in the future. The lock-up situation can be seen in apps for specific cloud platforms, such as Amazon EC2, Microsoft Azure, that are not easily transferred to any other cloud platform and that users are vulnerable to changes made by their providers to further confirm the lenses of a software developer. In fact, the issue of lock-in arises when, for example, a company decide to modify cloud providers (or perhaps integrate services from different providers), but cannot move applications or data across different cloud services, as the semantics of cloud providers' resources and services do not correspond. This heterogeneity of cloud semantics and APIs creates technological incompatibility which in turn leads to challenge interoperability and portability. This makes it very complicated and difficult to interoperate, cooperate, portability, handle and maintain data and services. For these reasons, from the point of view of the company it is important to maintain flexibility in changing providers according to business needs or even to maintain in-house certain components which are less critical to safety due to risks. The issue of supplier lock-in will prevent interoperability and portability between cloud providers. It is the way for cloud providers and clients to become more competitive.

10. Privacy and Legal issues

Apparently, the main problem regarding cloud privacy/data security is 'data breach.'

Infringement of data can be generically defined as loss of electronically encrypted personal information. An infringement of the information could lead to a multitude of losses both for the provider and for the customer; identity theft, debit/credit card fraud for the customer, loss of credibility, future prosecutions and so on. In the event of data infringement, American law requires notification of data infringements by affected persons. Nearly every State in the USA now needs to report data breaches to the affected persons. Problems arise when data are subject to several jurisdictions, and the laws on data privacy differ. For example, the Data Privacy Directive of the European Union explicitly states that 'data can only leave the EU if it goes to a 'additional level of security' country.' This rule, while simple to implement, limits movement of data and thus decreases data capacity. The EU's regulations can be enforced.

1.3 Historical developments

No state-of-the-art technology is cloud computing. The development of Cloud Computing through various phases, including Grid Computing, Utility Computing, Application Service Provision and Software as a Service, etc., has taken place. But the overall (whole) concept of the provision of computing resources via a global network began in the 1960s. By 2020, it is projected that the cloud computing market will exceed 241 billion dollars. But the history of cloud computing is how we got there and where all that started. Cloud computing has a history that is not that old, the first business and consumer cloud computing website was launched in 1999 (Salesforce.com and Google). Cloud computing is directly connected to Internet development and the development of corporate technology as cloud computing is the answer to the problem of how the Internet can improve corporate technology. Business technology has a rich and interesting background, almost as long as businesses themselves, but the development that has influenced Cloud computing most directly begins with the emergence of computers as suppliers of real business solutions.

History of Cloud Computing

Cloud computing is one of today's most breakthrough technology. Then there's a brief cloud-computing history.

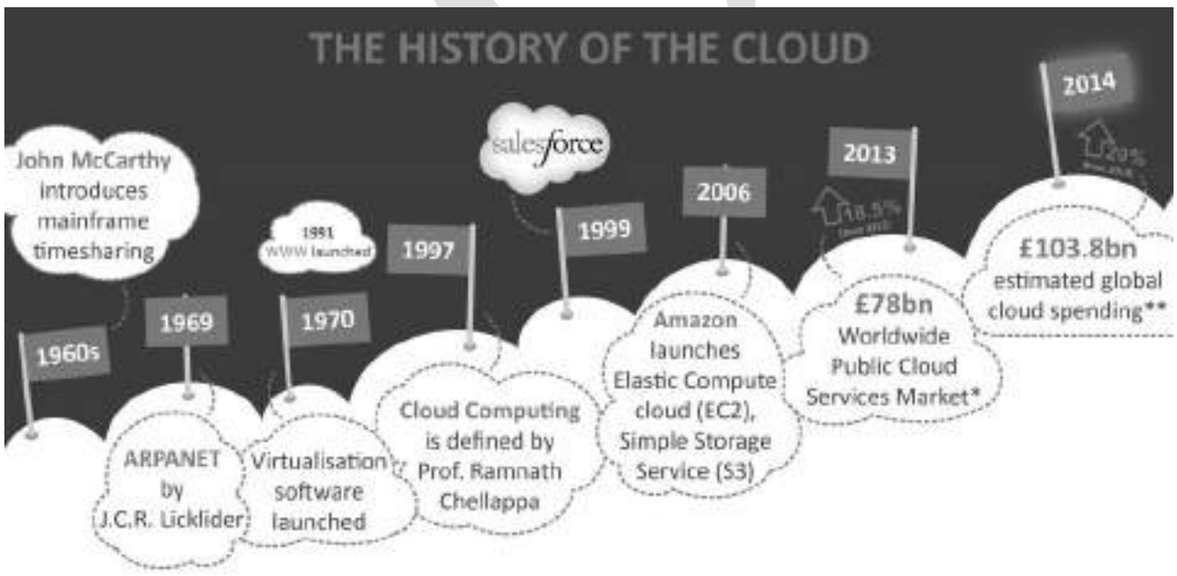


FIGURE 1. 9 History of Cloud Computing [*Gartner, **Constellation Research]

EARLY 1960S
Computer scientist John McCarthy has a time-sharing concept that allows the organization to use an expensive mainframe at the same time. This machine is described as a major contribution to Internet development, and as a leader in cloud computing.

IN 1969
J.C.R. Licklider, responsible for the creation of the Advanced Research Projects Agency (ARPANET), proposed the idea of an "Intergalactic Computer Network" or "Galactic Network" (a computer networking term similar to today's Internet). His vision was to connect everyone around the world and access programs and data from anywhere.

IN 1970
Usage of tools such as VMware for virtualization. More than one operating system can be run in a separate environment simultaneously. In a different operating system it was possible to

operate a completely different computer (virtual machine).

IN 1997
Prof Ramnath Chellappa in Dallas in 1997 seems to be the first known definition of "cloud computing," "a paradigm in which computing boundaries are defined solely on economic rather than technical limits alone."

IN 1999
Salesforce.com was launched in 1999 as the pioneer of delivering client applications through its simple website. The services firm has been able to provide applications via the Internet for both the specialist and mainstream software companies.

IN 2003
This first public release of Xen ,is a software system that enables multiple virtual guest operating systems to be run simultaneous on a single machine, which also known as the Virtual Machine Monitor (VMM) as a hypervisor.

IN 2006
The Amazon cloud service was launched in 2006. First, its Elastic Compute Cloud (EC2) allowed people to use their own cloud applications and to access computers. Simple Storage Service (S3) was then released. This incorporated the user-as-you-go model and has become the standard procedure for both users and the industry as a whole.

IN 2013
A total of £ 78 billion in the world 's market for public cloud services was increased by 18.5% in 2012, with IaaS as one of the fastest growing services on the market.

IN 2014
Global business spending for cloud-related technology and services is estimated to be £ 103.8 billion in 2014, up 20% from 2013 (Constellation Research).

Figure gives an analysis of the development of cloud computing distributed technologies. When we track the historic developments, we review briefly five key technologies that have played a significant role in cloud computing. They are distributed systems, virtualization, Web 2.0, service orientation and utility computing.

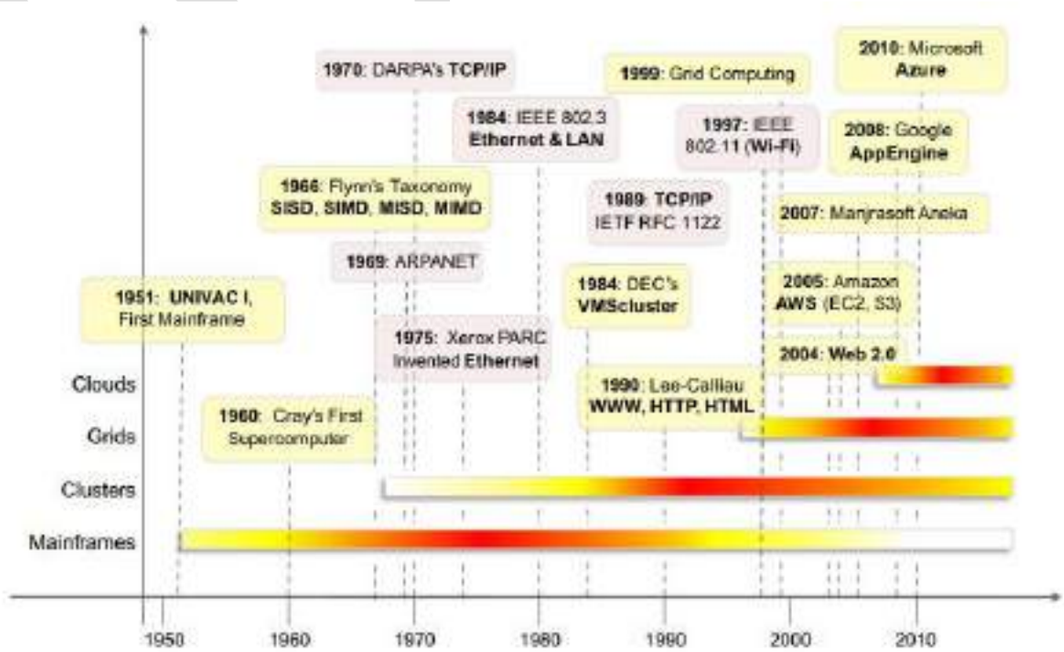


FIGURE 1.10: The evolution of distributed computing technologies, 1950s- 2010s.

Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

Distributed computing is a computer concept that refers most of the time to multiple computer systems that work on a single problem. A single problem in distributed computing is broken

down into many parts, and different computers solve each part. While the computers are interconnected, they can communicate to each other to resolve the problem. The computer functions as a single entity if done properly.

The ultimate goal of distributed computing is to improve the overall performance through cost-effective, transparent and secure connections between users and IT resources. It also ensures defect tolerance and provides access to resources in the event of failure of one component.

There really is nothing special about distributing resources in a computer network. This began with the use of mainframe terminals, then moved to minicomputers and is now possible in personal computers and client server architecture with several tiers.

A distributed computer architecture consists of a number of very lightweight client machines installed with one or several dedicated servers for computer management. Client agents normally recognize when the machine is idle, so that the management server is notified that the machine is not in use or that it is available. The agent then asks for a package. When this application package is delivered from the management server to the client, when it has free CPU cycles, the software runs the application software and returns the results to the management server. When the user returns, the management server will return the resources used to perform a number of tasks in the absence of the user.
Distributed systems show heterogeneity, openness, scalability, transparency, concurrency, continuous availability and independent failures. These characterize clouds to some extent, especially with regard to scalability, concurrency and continuous availability.
Cloud computing has contributed to three major milestones: mainframe, cluster computing and grid computing.

Mainframes: A mainframe is a powerful computer which often serves as the main data repository for an IT infrastructure of an organization. It is connected with users via less powerful devices like workstations or terminals. It is easier to manage, update and protect the integrity of data by centralizing data into a single mainframe repository. Mainframes are generally used for large-scale processes which require greater availability and safety than smaller machines. Mainframes computers or mainframes are primarily machines for essential purposes used by large organizations; bulk data processing, for example census, industry and consumer statistics, enterprise resource planning and transaction processing. During the late 1950s, mainframes only had a basic interactive interface, using punched cards, paper tape or magnetic tape for data transmission and programs. They worked in batch mode to support back office functions, like payroll and customer billing, mainly based on repetitive tape and merging operations followed by a line printing to continuous stationary pre-printed. Introducing digital user interfaces almost solely used to execute applications (e.g. airline booking) rather than to build the software. Typewriter and Teletype machines were standard network operators' control consoles in the early ' 70s, although largely replaced with keypads.

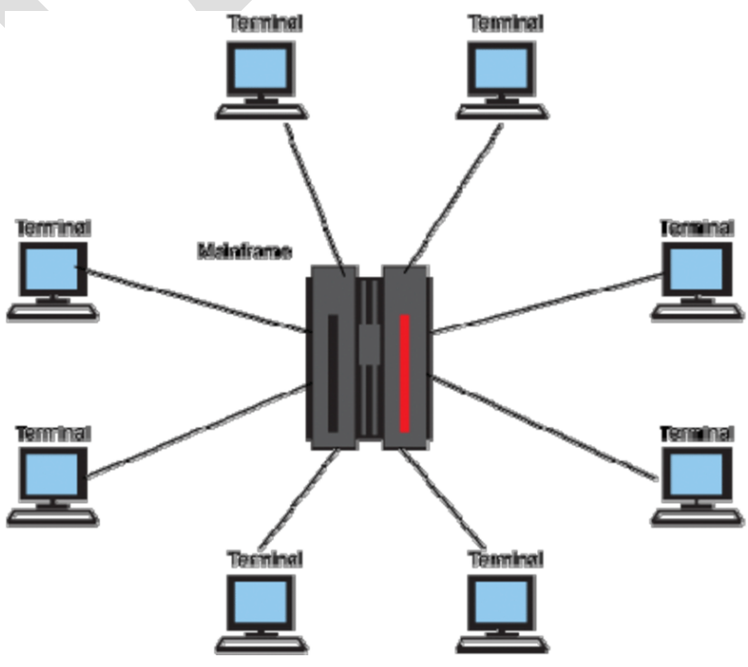


FIGURE 1.11 Mainframes

Cluster computing: The approach to computer clustering typically connects some computer

rental applications are delivered through networks. In essence, ASPs provided businesses with a way to outsource any or all parts of their IT needs.

The ASP maintains the responsibility for managing the application in its infrastructure, using the Internet as a connection between every customer and the key software application, through a centrally hosted Intent application. What this means for an organization is for the ASP to retention and guarantee the program and data are accessible whenever appropriate, including the related infrastructure and the customer data.

While the ASP model first introduced the software as a service definition, it was not able to provide full applications with customizable due to numerous inherent constraints such as its inability of designing extremely interactive applications.. The result has been the monolithic architectures and highly vulnerable integration of applications based on tight coupling principle in customer-specific architectures.
We are in the middle of yet another significant development today in the development of a software as a service architecture for asynchronous loosely linked interactions based on XML standards with the intention of making it easier for applications to access and communicate over the Internet. The SOC model enables the idea software-as-a-service to extend to use the provision of complicated business processes and transactions as a service, and allow applications to be created on the fly and services to be replicated across and by everyone. Many ASPs are pursuing more of a digital infrastructure and business models which are similar with those of cloud service providers to the relative advantages of internet technology.

Functional and non-functional attributes consist of the web services. Quality of service (QoS) is the so called unfunctional attributes. QoS is defined as a set of nonfunctional characteristics of entities used to move from a web service repository to consumers who rely on the ability of a web service to fulfill its specified or implied needs in an end-to - end way, according to the quality definition of ISO 8402. Examples of QoS features include performance, reliability, security, accessibility, usability, discovery, adaptively and composability. A SLA that identifies the minimum (or acceptable range) values for QoS attributes to be complied with on calling the service shall establish a QoS requirement between the clients and providers.

What is Service Oriented Architecture?
Service-oriented Architecture or SOA bring us all to understand it as a architecture which orients around services.. Services are discreet software components implemented using well-defined interface standards. Service is delivered to a directory or registry until it is created and validated to allow other developers to access the service. The registry also provides a repository that contains information on the published service, for example how to create the interface, what levels of service are required, how to retain authority, etc.

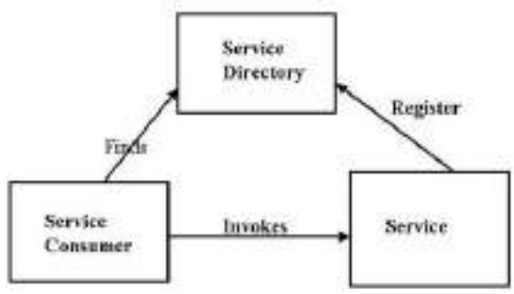


FIGURE 1.16 Service-oriented Architecture

SOA benefits
SOA services allow for agility of business. By integrating existing services, developers can create applications quickly.

The services are distinct entities and can be invoked without a platform or programming language knowledge at run-time.

The services follow a series of standards – Web Services Description Language (WSDL),

Representational State Transfer (REST), or the Simple Object Access Protocol(SOAP) – which facilitate their integration with both existing and new applications. The SOAP services are complemented by the following standards. SOAP.

Safety through Service Quality (QoS). Certain elements of QoS include authentication and authorisation, reliable and consistent messaging, permission policies, etc.
There is no interdependence of each other's service components.

SOA and cloud computing challenges
The network dependency of both of these technologies is one of the major challenges.
In addition, dependence on the cloud provider, contracts and service levels agreements is the challenges specific to cloud computing.

One of the challenges for SOA today are the requests to improve or change the service provided by SOA service providers.

Does Cloud Computing compete with SOA?

Some see cloud computing as a descendant of SOA. It would not be completely untrue, as the principles of service guidelines both apply to cloud computing and SOA. The following illustration shows how Cloud Computing Services overlap SOA-

Cloud Computing	Overlap	SOA via Web Services
• Software as a Service (SaaS) • Utility Computing • Terabytes on Demand • Data Distributed in a Cloud • Platform as a Service • Standards Evolving for Different Layers of the Stack	• Application Layer Components/Services • Network Dependence • Cloud/IP Wide Area Network (WAN)-supported Service Invocations • Leveraging Distributed Software Assets • Producer/Consumer Model	• System of Systems Integration Focus • Driving Consistency of Integration • Enterprise Application Integration (EAI) • Reasonably Mature Implementing Standards (REST,SOAP,WSDL, UDDI,etc.)

It is very important to realize that while cloud computing overlaps with SOA, they focus on various implementation projects. In order to exchange information between systems and a network of systems, SOA implementations are primarily used. Cloud computing, on the other hand, aims to leverage the network across the whole range of IT functions.

SOA is not suitable for cloud computing, actually they are additional activities. Providers need a very good service-oriented architecture to be able to provide cloud services effectively.

There are many common features of SOA and cloud computing, however, they are not and can coexist. In its requirements for delivery of digital services, SOA seems to have matured. Cloud Computing and its services are new as are numerous vendors such as public, community, hybrid and private clouds, with their offerings. They are also growing.

1.3.4 Utility-oriented computing

The concept Utility Computing pertains to utilities and business models that provide its customers with a service provider, and charges you for consumption. The computing power, storage or applications are examples of such IT services. In this scenario the customer will be the single divisions of the company as a service provider at a data center of the company.

The concept utility applies to utility services offered by a utilities provider, such as electricity, telephone, water and gas. Related to electricity or telephone, where the consumer receives the utility computing, computing power is measured and paid on the basis of a shared computer network.

The concept utility applies to utility services offered by a utilities provider, such as electricity, telephone, water and gas. Related to electricity or telephone, where the consumer receives the utility computing, computing power is measured and paid on the basis of a shared computer network.

Utility computing is very analogous to virtualization so that the total volume of web storage and the computing capacity available to customers is much greater than that of a single computer. To make this type of web server possible, several network backend servers are often used. The dedicated webservers can be used in explicitly built and leased cluster types for end users. The distributed computing is the approach used for a single 'calculation' on multiple web servers.

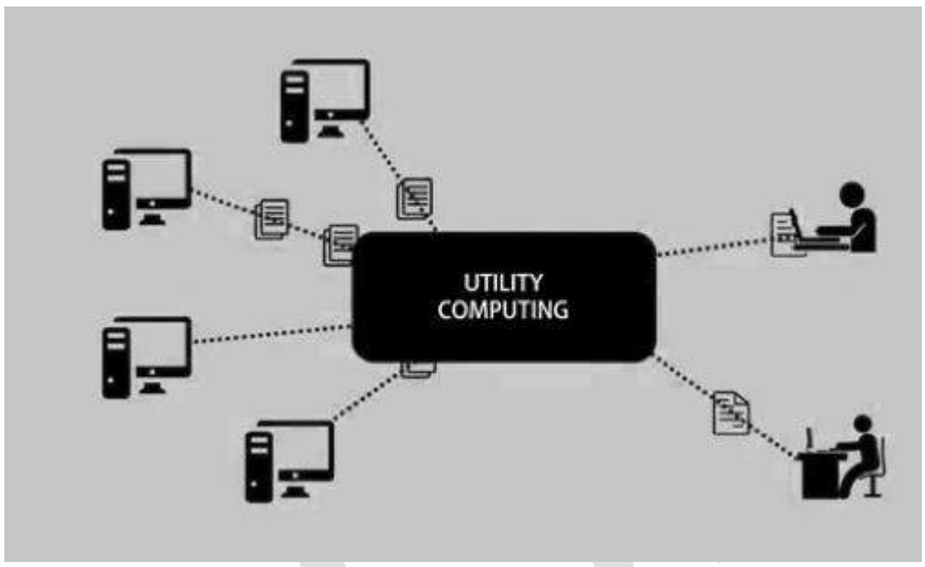


FIGURE 1.17 Cloud Computing Technology – Utility Computing

Properties of utility computing

Even though meanings of utility computing are various, they usually contain the following five characteristics.

Scalability

The utility computing shall ensure that adequate IT resources are available under all situations. Improved service demand does not suffer from its quality (e.g. response time).

Price of demand

Until now, companies must purchase their own computing power such as hardware and software. It is necessary to pay for this IT infrastructure beforehand, irrespective of its use in the future. For instance, technology providers to reach this link depends on how many CPUs the client has enabled during leasing rate for their servers. If the computer capacity to assert the individual sections actually can be measured in a company, the IT costs may be primarily attributable to each individual unit at internal cost. Additional forms of connection are possible with the use of IT costs.

Standardized Utility Computing Services

A collection of standardized services is accessible from the utility computing service provider. These agreements may differ in the level of service (Quality Agreement and IT Price). The consumer does not have any impact on the infrastructure, such as the server platform

Utility Computing and Virtualization

Virtualization technologies can be used to share web and other resources in the common pool of machines. Instead of the physical resources available, this divides the network into logical resources. No predetermined servers or storage of any other than a free server or pool memory are assigned to an application.

Automation

Unit 1

Chapter 2

Unit Structure

2.0 Objective

The computing components (hardware, software, infrastructures) that allow the delivery of cloud computing services refer to a Cloud system or cloud computing technology.

Consumers can acquire new skills without investing in new hardware or software via the public cloud. Instead, they pay a subscription fee for their cloud provider or only pay for their resources. These IT assets are owned and managed through the Internet by the service providers.

This chapter presents the basic principles and models of parallel and distributed computing, which provide the foundation for building cloud computing systems and frameworks.

2.1 Eras of computing

The two most prominent computing era are sequential and parallel. In the past decade, the high performance computer searches for parallel machines have become important competitors of vector machines. Figure 2.1 provides a hundred-year overview of the development of the computing era. During these periods the four main computing elements are created like architectures, compilers, applications and problem-solving environments.

The computing era begins with the development of hardware, followed by software systems (especially in the area of compilers and operating systems), applications, and with a growing problem solving environment it enters its saturation level. Each computing element is subject to three stages: R&D, commercialization and commodity.

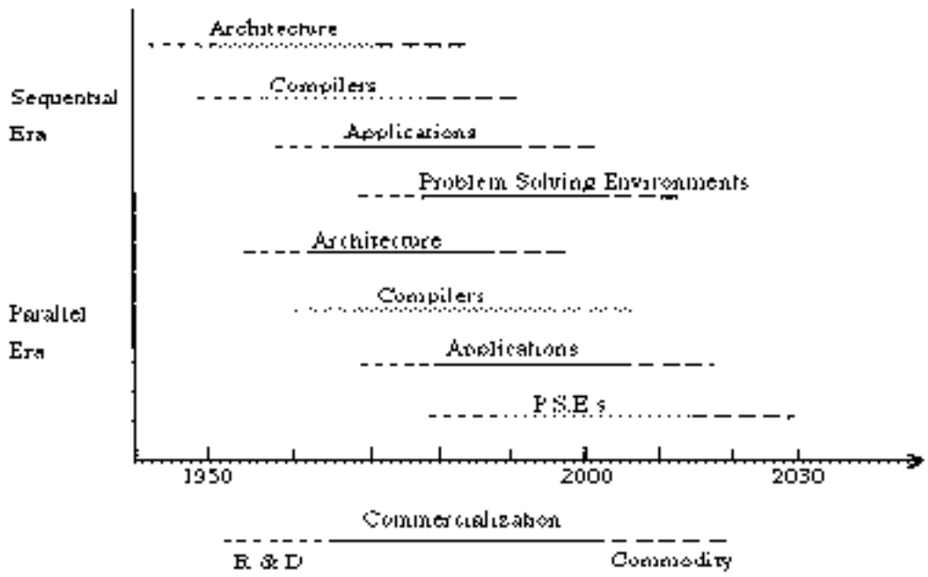


FIGURE 2.1 Eras of computing

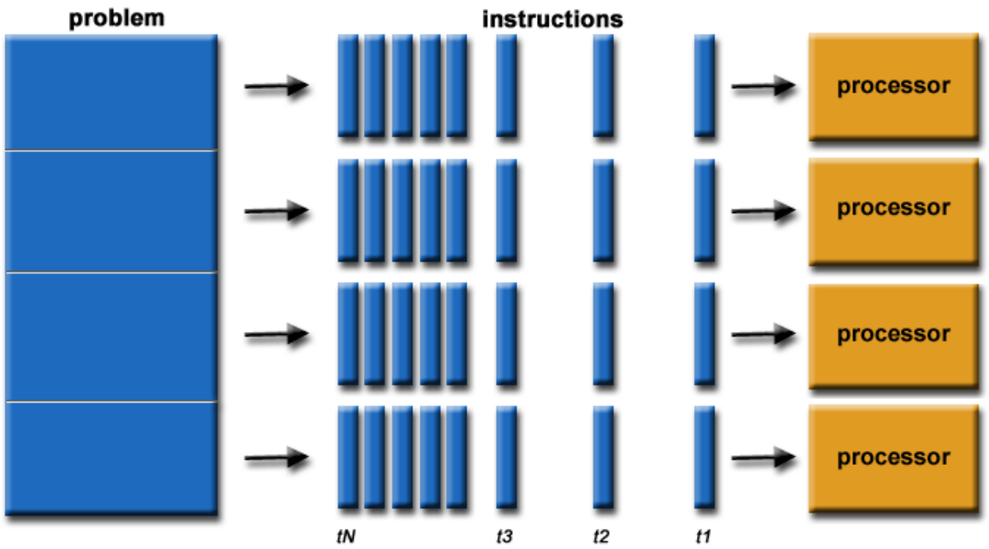
2.2 Parallel vs. distributed computing

Parallel computing and distributed computing terms, even though they are somewhat different things, are often used interchangeably. The term parallel means a tightly coupled system, whereas the distributed one refers to a wider system class, including the tightly coupled.

The concurrent use of several computer resources to solve a computational problem is parallel computing:

- A problem is divided into discrete pieces which can be solved simultaneously
- A number of instructions for each part are broken down further

- Instructions on various processors from each part run simultaneously
- An overall mechanism for control/coordination is used



For example:

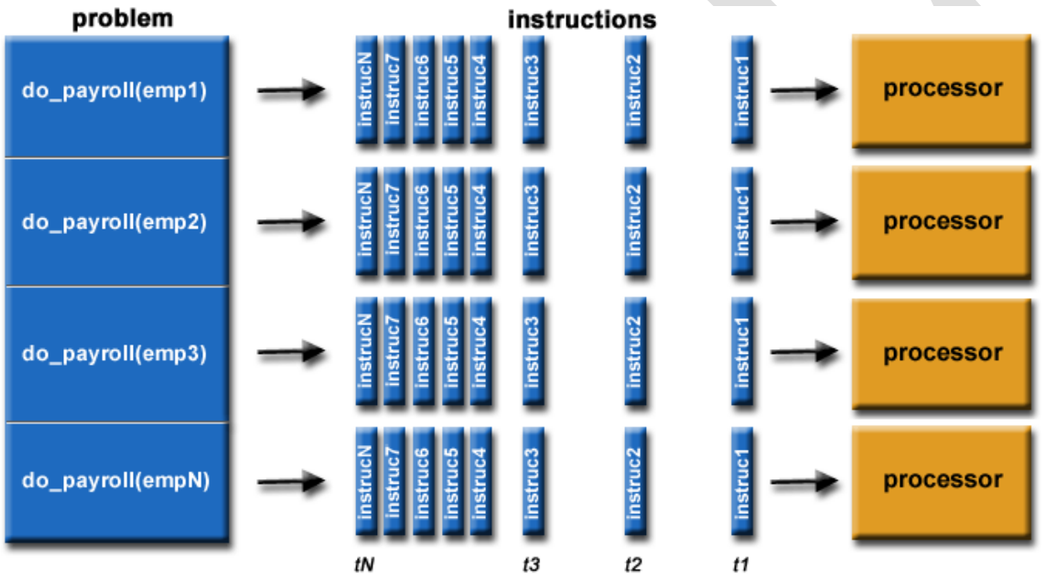


FIGURE 2.2 Sequential and Parallel Processing

The problem in the computation should be:

The problem in the computation should be:

- Be divided into discreet parts of work that can be solved at the same time;;
- At any given time, execute multiple program instructions;
- With many compute resources in less time than one compute resource, can be solved.

Typically, computation resources are:

- One computer with several processors / cores
- A random number of these computers connected through a network

Initially, only certain architectures were considered by parallel systems .It featured multiple processors with the same physical memory and a single computer. Over time, those limitations have been relaxed and parallel systems now include all architectures, whether physically present or based on the concept of shared memory, whether the library support, specific hardware and a very efficient network infrastructure are present physically or created. For example, a cluster of the nodes linked by InfiniBand can be considered a parallel system and configured with a distributing shared memory system

This paradigm separates the implementation of the component from the knowledge of component names and locations. The pattern of the publisher / subscriber, where:

Publisher(s): advertise the data you would like to share with others
Subscriber(s): Receipt of published data register interest.

For communications between components, a message manager is used. Publishers send messages to the manager who redistributes them to subscribers.

Communication process
The architectural type of communication process is also known as Client-Server architecture.
Client: begins a server call that requests for some service.
Server: provides client data.
Returns data access when the server works synchronously

2.4.3.3 System architectural styles

The Client-server and Peer-to - peer (P2P) are the two key system level architectures we use today. In our everyday lives, we use these two types of services, but the difference between them is often misinterpreted.

Client Server Architecture

Two major components are in the client server architecture. The server and the client. The server is the location of all transmission, transmission, and processing data, while the client can access the remote server services and resources. The server allows clients to make requests, and the server will reply. In general, the remote side is managed only by a computer. But in order to be on the safe side, we load balancing techniques using several servers.

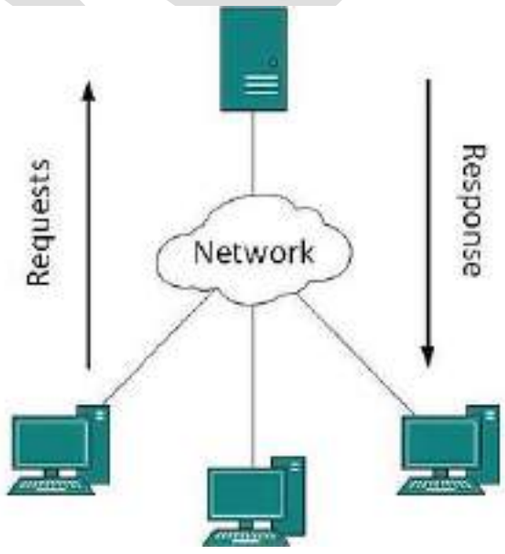


FIGURE 2.18 Client/server architectural styles

The Client Server architecture is a standard design feature with a centralized security database. This database includes information on security, such as credentials and access details. Absent security keys, users can't sign in to a server. This architecture therefore becomes a bit more stable and secure than Peer to Peer. The stability comes because the security database can make for more efficient use of resources. However, on the other hand, the system could crash because only a small amount of work can be done by a server at a certain time.

Advantages:

- Easier to Build and Maintain

- Better Security
- Stable

Disadvantages:

- Single point of failure
- Less scalable

Peer to Peer (P2P)

There is no central control in a distributed system behind peer to peer. The fundamental idea is that at a certain time each node can be a client or a server. If something is asked from the node, it could be referred to as a client and if something arrives from a node it could be referred to as a server. Usually every node is called a peer.

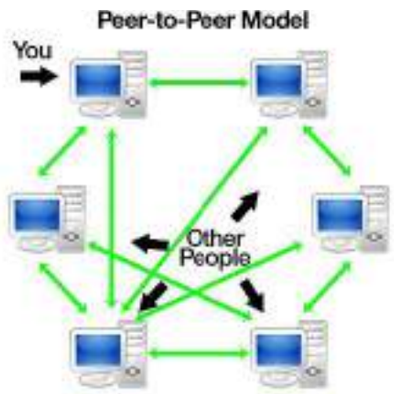


FIGURE 2.19 Peer to Peer (P2P)

Any new node will first join this network. Upon joining, they may either request or provide a service. A node's initiation phase (joining a node) can vary based on network's implementation. There are two ways a new node can learn what other nodes provide.

Centralized Lookup Server-The new node must register and mention the services on the network with the centralized look up server. So, just contact the centralized look up system anytime you need to have a service and it will direct you to the appropriate service provider.

Decentralized System-A node that seeks particular services will, broadcast and request each other node in the network, so that the service provider can respond.

A Comparison between Client Server and Peer to Peer Architectures

BASIS FOR COMAPARISON	CLIENT-SERVER	PEER-TO-PEER
Basic	There is a specific server and specific clients connected to the server	Clients and server are not distinguished; each node act as client and server.
Service	The client request for service and server respond with the service.	Each node can request for services and can also provide the services.
Focus	Sharing the information.	Connectivity.
Data	The data is stored in a centralized server.	Each peer has its own data.
Server	When several clients request for the services simultaneously, a server can get bottlenecked.	As the services are provided by several servers distributed in the peer-to-peer system, a server in not bottlenecked.
Expense	The client-server are expensive to implement.	Peer-to-peer are less expensive to implement.

- The code cache is used to store the most used translated instructions for improving performance, but it increases memory usage with hardware costs.
- On the x86 architecture, the performance of the complete virtualization is 80-95% of the host machine.

3.13.5 Virtualization solutions

VMware is a pioneer in virtualization and cloud infrastructure solutions that allow our 350,000-plus enterprise and customers to succeed in the cloud age. VMware simplifies its complexity across the entire data center and enables customers with software-defined data center solutions to become hybrid cloud computing and the mobile workspace.

3.13.5 End-user (desktop) virtualization

VMware desktop and app-virtualization technologies give IT a streamlined method for providing, securing and maintaining Windows and Linux desktops and applications on site or on the cloud, reducing costs and ensuring end users can operate everywhere and everywhere. VMware Workstation allows users to run different operating systems on one and the same Windows or Linux PC concurrently. Create real Linux and Windows VMs as well as other desktop, server and tablet environments, complete with virtual networking configurable and network simulation, for use with code development, solution architecture and application testing and product demonstrations, and much.
VMware Fusion allows Mac users the ability to run Windows on Mac together with hundreds of several other operating systems side-by - side without rebooting. Fusion is easy enough for home users and sufficiently efficient for IT experts, developers and businesses. Other than setting up an independent computing environment, the two products enable a guest operating system to exploit host machine resources (USB devices, folder sharing and integration with the host operating system's graphical user interface (GUI). Figure provides a description of the systems' architecture.

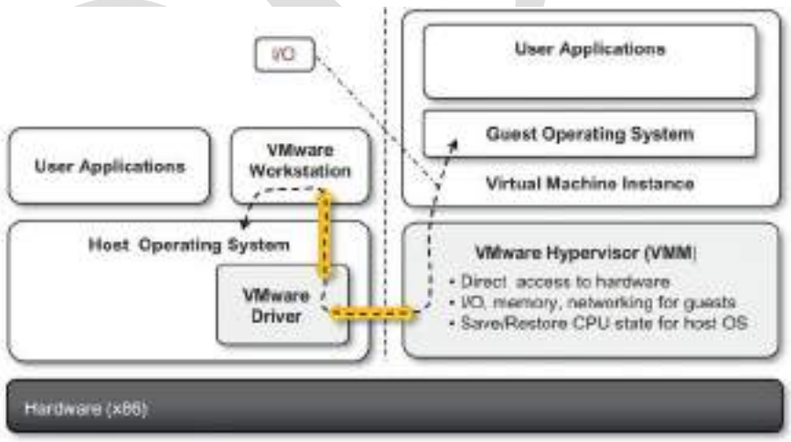


FIGURE 3.18 VMware workstation architecture.
(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

A guest operating system installed application creates the virtualization environment, which enables those operating systems to fully virtualize the hardware that underlie it. This is done through the installation in the host operating system of a special driver which provides two main services:

- It uses a virtual machine manager, which can be used in privileged mode.

- The VMware application offers straps for processing particular I / O requests by subsequently forwarding these requests via system calls to the host operating system.

This architecture, also known as Hosted Virtual Machine Architecture, can both separate virtual machine instances inside an application's memory space and provide decent efficiency, as VMware application's involvement is only essential for instructions, for example, I / O devices which require binary translation. The Virtual Machine manager manages the CPU and the MMU and changes the operational functionality of the CPU and MMU with the host OS. Virtual machine images are stored in a host file system catalogue, and that both VMware and VMware Fusion allow new images to be created, run, create snapshot and undo operational activities by turning back to a previous virtual machine state

VMware Player, VMware ACE and VMware ThinApp are additional technologies relevant to the virtualization of end-user computing environments. VMware Player is a limited VMware Workstation version which enables the creation and emulation of virtual machines of an operating environment such as Windows or Linux. VMware ACE is same as VMware Workstation for developing the policy wrapped virtual machines for provisioning the secure deployment of client virtual environments on end user computers. VMware ThinApp is an application's virtualization solution. It offers an independent development environment to prevent variations due to versioning and incompatible applications. It identifies the operating environment changes by installing a specific app and stores these in a package that can be executed with VMware ThinApp along with the binary app.

3.13.6 Server virtualization

GSX Server is a Windows and Linux virtualized server system developed and distributed by VMware, a subsidiary of EMC Corporation, The program promotes remote management, provisioning and application standardization. Figure demonstrates the architecture of the VMware GSX Server.

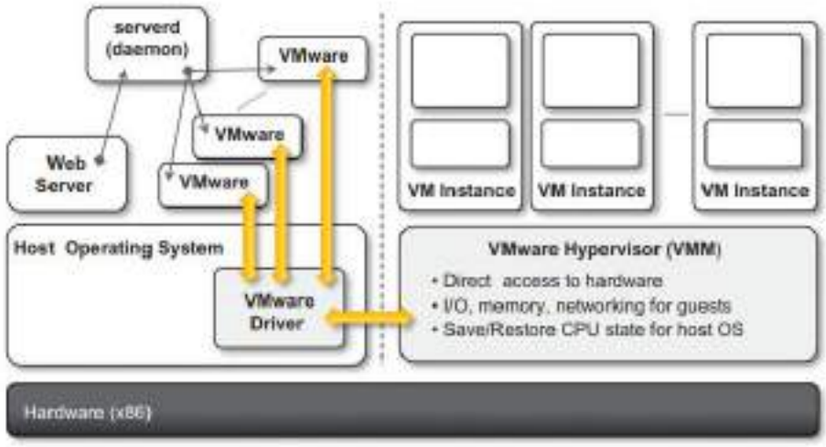


FIGURE 3.19 VMware GSX server architecture.
(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

VMware GSX server converts computers into a collection of virtual machines. Operating systems and frameworks are in separation different Virtual machines on a single hardware device. VMware GSX Server offers wide support for inherited hardware support for the device from the host. The reliable architecture and integration capability of the product VMware GSX Server makes Windows and Linux host environments simple to use and manage. A host program for VMware GSX Server helps you to deploy, monitor and control your application and multiple servers on virtual machines operating remotely
The architecture is designed primarily for web server virtualization. A serverd called daemon process monitors and manages applications for VMware. The VMware driver on the host operating system then connects these programs to the virtual machine instances. The VMM is used to handle virtual machine instances as earlier defined. User requests for managing and providing virtual machines are redirected from the Web server via the serverd via the VMM.

The hypervisor-based method is demonstrated by VMware ESX Server and its improved edition VMWare ESXi Server. Each can be mounted on bare metal servers and provide VM management services. These two products offer similar services, but differ in the internal architecture, particularly in the hypervisor kernel organization. An updated Linux operating system version that allows access to the hypervisor via the service console is implemented by VMware ESX. VMware ESXI introduces an incredibly thin OS layer and substitutes the service console with remote monitoring interfaces and utilities, thus greatly reducing hypervisor code size and memory footprint Figure demonstrates the architecture of VMware ESXi.

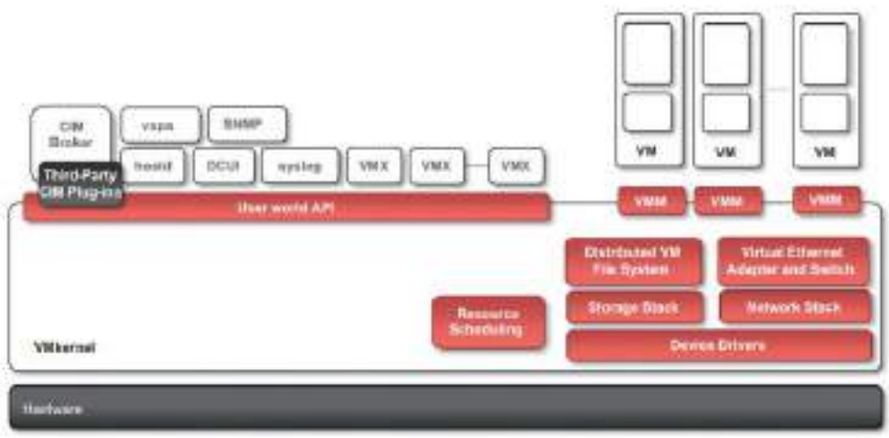


FIGURE 3.20 VMware ESXi server architecture
(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

3.14 Microsoft Hyper-V

Hyper-V is a hypervisor, a virtualizing software of Microsoft which can build and host the enterprise virtual machines on x86-64 systems, such as desktops and servers. Basically, hypervisor is a software that permits several virtual servers (guest machine) to be operated on the physical (host) server. A Hyper-V server may be configured to expose virtual machines to one or more networks. The Windows Server 2008 first launched Hyper-V

3.14.1 Architecture

In terms of partition, Hyper-V implements virtual machine isolation. A partition, supported by the hypervisor, is a logical isolation unit in which every guest operating device performs. In a hypervisor instance, there must be at least one parent partition running Windows Server (2008 and later) enabled. Within the parent partition the virtualization software works and has direct access to hardware devices. The parent partition produces child's partitions containing the guest operating systems. A parent partition uses a hypercall API, the application framework, to build children's partitions exposed to the Hyper-V.

A child partition has no access or its actual interrupts to the physical processor. Rather, the processor has a virtual view and operates in the guest virtual address, which may not actually be the entire virtual address space depending upon on configuration of the hypervisor. Hyper-V will only show a subset of processors on each partition, depending on the VM configuration. The hypervisor manages the interrupts to the processor with the aid of a logical Synthetic Interrupt controller (SynIC) to the respective partition.

Child partitions provide a virtual view of the resources of the virtual machines, rather than a direct access for hardware sources. Each virtual device request is sent to the parent partition devices, which handle the requests, through the VMBus. The VMBus is a logical channel that allows communication between sections. The response is also forwarded through the VMBus. When the parent partition devices are also virtual devices, they are routed to the parent partition where physical devices are accessed. Parent partitions operate a Virtualization Service Provider (VSP) that connects to the VMBus and handles requests for system access from the child partitions. Child's virtual partition devices run the

Virtualization Service Client (VSC) internally, redirecting the program to the VSPs on the VMBus parent. To the guest OS, this whole process is transparent.

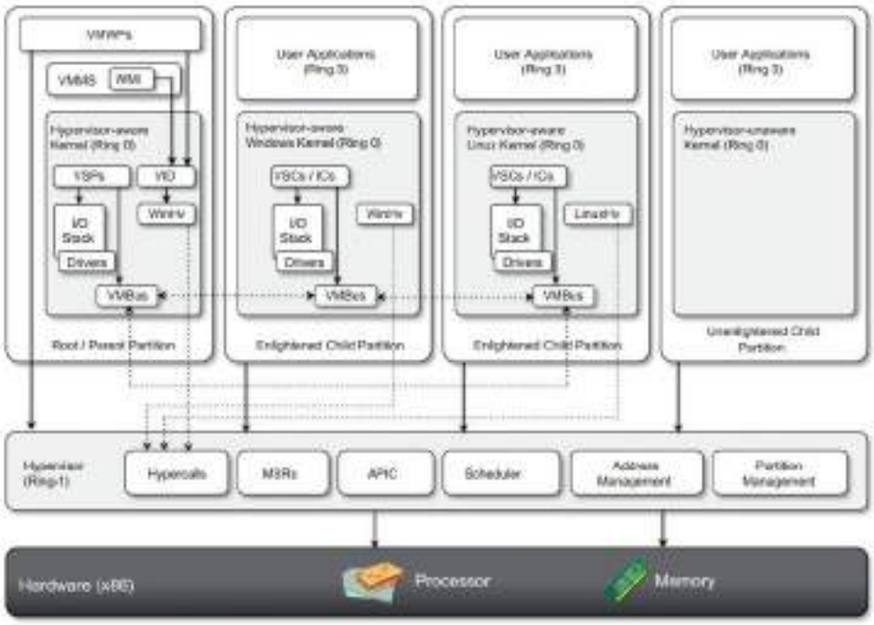


FIGURE 3.17 Microsoft Hyper-V architecture

(Reference from "Mastering Cloud Computing Foundations and Applications Programming" by Rajkumar Buyya)

3.15 Summary

Virtualization is an essential framework for a number of technologies and concepts. The popular source for virtualization is the ability to demonstrate, using some kind of emulation or abstraction layer, a given runtime environment, whether a software, a storage facility, a network connection or a remote desktop. In building cloud services and infrastructure, all these principles play an essential role, wherein hardware, IT infrastructure, applications and services are provided on demand via the Internet or usually through a network connection.

3.16 Review questions

1. Define Virtualization. What are the advantages of virtualization?
2. What are the characteristics of virtualized environments?
3. Describe classification or taxonomy of virtualization at different levels.
4. Discuss the execution virtualization machine reference model.
5. What are the techniques of hardware virtualization?
6. List and discuss various virtualization types.
7. What are the benefits of virtualization in the context of cloud computing?
8. What are the disadvantages of virtualization?
9. What is Xen? Discuss its elements for virtualization.
10. Discuss the reference model of full virtualization.
11. Discuss the reference model of paravirtualization.
12. Discuss the architecture of Hyper-V. Discuss its use in cloud computing

3.17 Reference for further reading

1. Mastering Cloud Computing Foundations and Applications Programming Rajkumar Buyya ,Christian Vecchiola,S. Thamarai Selvi MK publications ISBN: 978-0-12-411454-8

2. Cloud Computing Concepts, Technology & Architecture Thomas Erl, Zaigham Mahmood, and Ricardo Puttini , The Prentice Hall Service Technology Series ISBN-10 : 9780133387520 ISBN-13 : 978-0133387520

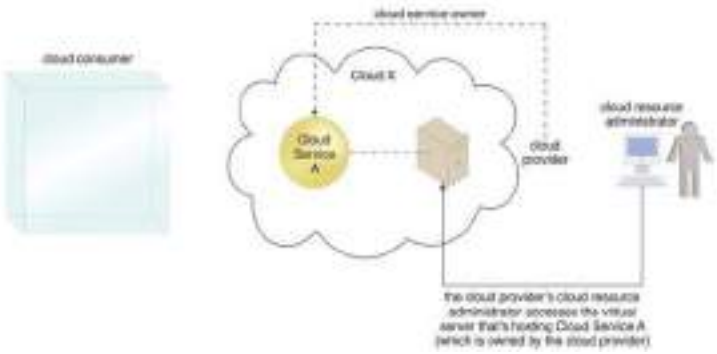


Figure2.1.5. A cloud resource administrator can be with a cloud provider organization for which it can administer the cloud provider's internally and externally available IT resources.

(Reference : Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini)

The explanation a cloud asset executive isn't alluded to as a "cloud administration manager" is on the lands that this job might be answerable for directing cloud-based IT assets that don't exist as cloud administrations. For instance, if the cloud asset chairman fits to (or is Shrunk by) the cloud supplier, IT assets not made remotely available might be controlled by this job (and these sorts of IT assets are not delegated cloud administrations).

2.1.4. Limit

2.1.4.1. Hierarchical Boundary

A hierarchical limit implies the physical outskirt that conditions a lot of IT capitals that are had and controlled by an association. The authoritative limit doesn't show the limit of a real association, just a hierarchical arrangement of IT properties and IT assets. Also, mists have an authoritative limit (Figure 2.1.6.).

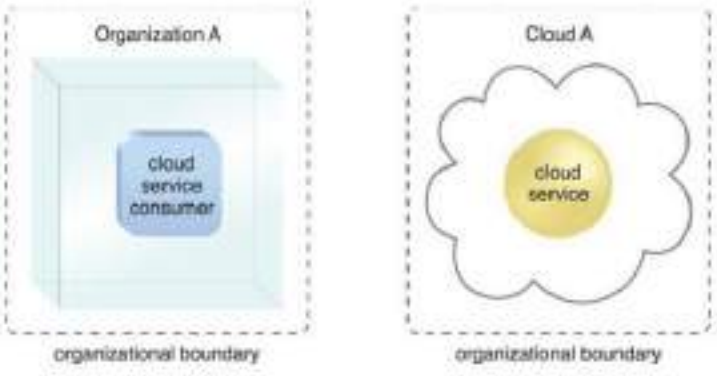


Figure2.1.6. Organizational boundaries of a cloud consumer (left), and a cloud provider (right), represented by a broken line notation.

(Reference : Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini)

2.1.4.2. Trust Boundary

At the point when an association embraces the job of cloud client to get to cloud-based IT assets, it needs to spread its trust outside the physical limit of the association to contain portions of the cloud condition. A trust limit is a coherent fringe that normally ranges outside physical limits to connote the degree to which IT assets are trusted (Figure 2.1.7). When investigating cloud situations, the trust limit is most every now and again associated with the trust gave by the association going about as the cloud purchaser.

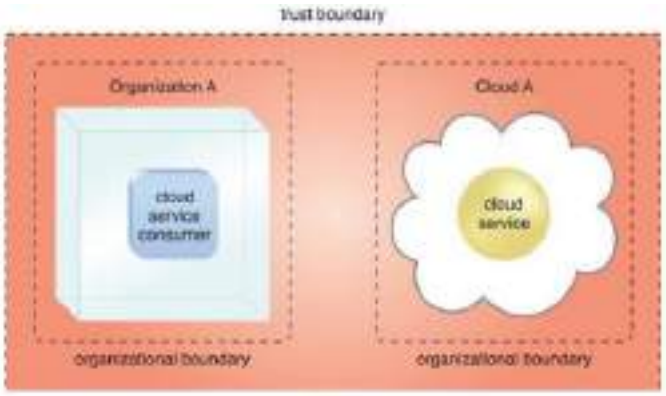


Figure 2.1.7. An extended trust boundary encompasses the organizational boundaries of the cloud provider and the cloud consumer.

(Reference : Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini)

2.1.5. Cloud Characteristics

An IT air needs a careful arrangement of qualities to empower the removed provisioning of walkable and estimated IT capitals in a real manner. These qualities need to exist to an expressive degree for the IT climate to be estimated a viable cloud.

The accompanying six accurate attributes are normal to the standard of cloud situations: • on-request utilization
• pervasive access
• multitenancy (and asset pooling)
• flexibility
• estimated utilization
• strength
Cloud suppliers and cloud clients can quantify these attributes independently and together to gauge the worth commitment of a given cloud stage. Despite the fact that cloud-based administrations and IT assets will get and display particular qualities to variable degrees, generally the better how much they are strengthened and applied, the better the resulting esteem proposal.

2.1.5.1. On-Demand Usage

A cloud client can separately get to cloud-based IT properties giving the cloud client the self-rule to self-arrangement these IT properties. When sorted out, utilization of oneself provisioned IT properties can be programmed, needful no extra human interest by the cloud client or cloud supplier. This outcomes in an on-request utilization circumstance. Otherwise called "on-request self-administration utilization," this trademark permits the administration based and use driven highlights begin in customary mists.

2.1.5.2. Universal Access

Universal access connotes the fitness for a cloud administration to be widely accessible. Establishing omnipresent access for a cloud administration can require arrangement for a scope of systems, transportation conventions, limits, and wellbeing advances. To allow this degree of access normally needs that the cloud administration design be customized to the particular prerequisites of various cloud administration clients.

2.1.5.2.1. Multitenancy (and Resource Pooling)

The quality of a product bundle that permits a case of the program to support various clients (occupants) whereby each is remote from the other, is referenced to as multitenancy. A cloud supplier pools its IT properties to enable various cloud to support clients by utilizing multitenancy imitations that routinely trust on the utilization of virtualization advancements. Over the utilization of multitenancy innovation, IT properties can be animatedly allotted and reallocated, rendering to cloud administration client requests.

2.1.5.4. Versatility

Versatility is the programmed inclination of a cloud to unmistakably scale IT properties, as required in answer to runtime circumstances or as customized by the cloud client or cloud supplier. Versatility is regularly estimated a center resistance for the acknowledgment of distributed computing, primarily because of the way that it is firmly related with the Abridged Asset and Comparative Costs advantage. Cloud suppliers with tremendous IT properties can offer the most noteworthy scope of versatility

2.1.5.5. Estimated Usage

The deliberate utilization trademark means the fitness of a cloud stage to keep way of the use of its IT assets, for the most part by cloud clients. Established on what is estimated, the cloud supplier can charge a cloud client just for the IT properties truly utilized as well as for the time span through which access to the IT properties was chosen. In this unique circumstance, estimated utilization is firmly associated with the on-request trademark.

2.1.5.6. Strength

Strong computing is a type of failover that dispenses excess utilizations of IT properties across physical spots. IT properties can be pre-arranged so that in the event that one gets lacking, agreement is consequently given over to extra excess application. Inside distributed computing, the attribute of flexibility can make reference to repetitive IT properties inside a similar cloud (however in various physical areas) or over various mists. Cloud clients can

development both the reliability and availability of their applications by utilizing the strength of cloud-based IT properties.

2.1.6. Cloud conveyance model

A cloud conveyance model connotes an assigned, pre-bundled blend of IT assets available by a cloud supplier. Three common cloud conveyance models turned out to be comprehensively perceived and honorable:

• Infrastructure-as-a-Service (IaaS)

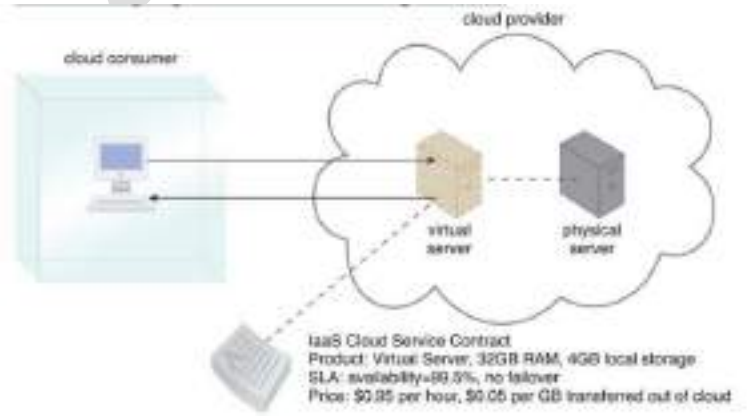
• Platform-as-a-Service (PaaS)

• Software-as-a-Service (SaaS)

2.1.6.1. Framework as-a-Service (IaaS)

The IaaS circulation model implies an independent IT climate contained of foundation driven IT assets which will be recovered and accomplished by means of cloud administration based interfaces and devices. This climate can incorporate equipment, organize, network, working frameworks, and other "crude" IT assets. In distinction to conventional facilitating or redistributing environmental factors, with IaaS, IT assets are normally virtualized and pressed into wraps that compress in advance runtime climbing and customization of the framework. the broadly useful of an IaaS domain is to flexibly cloud customers with an elevated level of control and responsibility over its.

design and use. The IT assets gave by IaaS are by and large not pre-arranged, setting the official obligation straightforwardly upon the cloud shopper. This model is consequently utilized by cloud buyers that need a significant level of command over the cloud-based condition they will make. Here and there cloud suppliers will contract IaaS contributions from other cloud suppliers in order to scale their own cloud surroundings. the sorts and makes of the IT assets gave by IaaS items offered by various cloud suppliers can change. IT assets accessible through IaaS conditions are for the most part offered as newly instated virtual occurrences. A focal and first IT asset inside a run of the mill IaaS condition is that the virtual server. Virtual servers are rented by indicating server equipment necessities, similar to processor limit, memory, and local space for putting away, as appeared in Figure



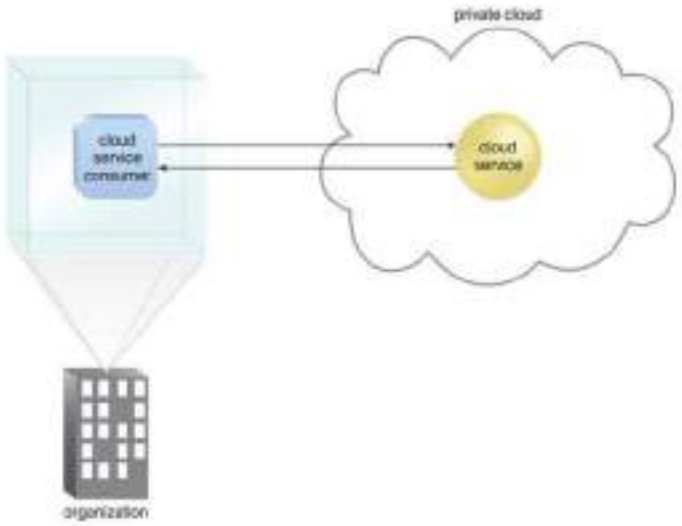


Figure 2.1.12.1. A cloud service consumer within the organization's on-premise environment accesses a cloud service hosted on an equivalent organization's private cloud via a virtual private network.

(Reference : Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini)

It is critical to utilize the affinities "on-reason" and "cloud-based" accurately inside the setting of an individual cloud. but the private cloud may genuinely dwell on the association's premises, IT assets it armed forces are as yet estimated "cloud-based" insofar as they're made remotely open to cloud shoppers. IT assets facilitated outside of the private cloud by the segments going about as cloud clients are along these lines considered "on-premise" regarding the private cloud-based IT assets.

2.1.7.4. Half and half Clouds

A half and half cloud may be a cloud environment included of at least two differing cloud arrangement models. for example , a cloud client may like to carry cloud managements making touchy information to an individual cloud and extra, fewer delicate cloud administrations to an open cloud. The aftereffects of this blend might be a half and half organization model (Figure 2.1.12.1).

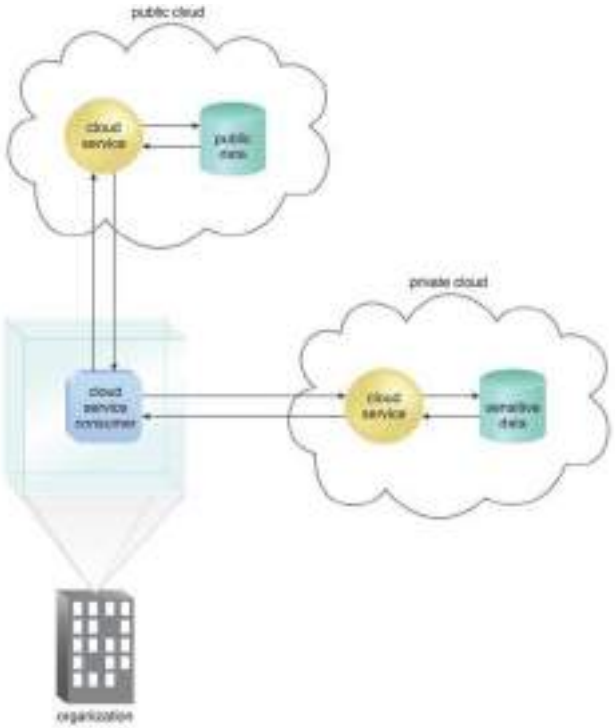


Figure 2.1.14. a establishment employing a hybrid cloud architecture that uses both a individual

and public cloud.

(Reference : Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini)

Mixture mien structures are regularly compound and testing to make and proceed with gratitude to the conceivable contrast in cloud airs and in this way, the undeniable actuality that organization accountabilities are traditionally part between the private cloud supplier association and the open cloud supplier.

2.1.8. Financial matters of the cloud

The principle drivers of distributed computing are economy of scale and simplicity of programming conveyance and its activity. Indeed, the most significant favorable position of this wonder is monetary: the pay-more only as costs arise model offered by cloud suppliers. particularly, distributed computing permits:

• Reducing the capital expenses related with the IT framework

• Eliminating the devaluation or lifetime costs identified with IT capital resources

• Replacing programming permitting with memberships

• Cutting the upkeep and regulatory expenses of IT assets

An expense of capital is that the expense happened in buying an advantage that is valuable inside the creation of items or the rendering of administrations. Capital expenses are one-time

costs that are commonly paid forthright which will contribute over the future to get benefit. The IT foundation and along these lines the product is capital resources since endeavors expect them to lead their business. at the present, it doesn't make a difference whether the foremost business of an endeavor is said there to in bright of the fact that the business will surely have an IT office that is wont to robotize a considerable lot of the exercises that are performed inside the venture: finance, client relationship the board, undertaking asset arranging, following and stock of items, and others. Subsequently, IT assets comprise an expense of capital for any very endeavor. it's acceptable practice to embrace to remain capital costs low since they present costs which will create benefit after some time; very that, since they're identified with material things they're dependent upon devaluation above the extensive run, which inside the end diminishes the benefit of the undertaking in light of the fact that such expenses are legitimately deducted from the venture incomes. inside the instance of IT capital costs, the deterioration costs are spoken to by the less valuable of the equipment after some time and in this manner the maturing of programming items that require to get supplanted on the grounds that new highlights are required.

Before distributed computing diffused inside the venture, the financial plan spent consequently framework and programming comprised a major cost for medium-sized and gigantic endeavors. Numerous undertakings own a little or medium-sized datacenter that presents a few operational expenses as far as upkeep, power, and cooling. Extra operational expenses are happened in keeping up an IT division and an IT bolster focus. In addition, different expenses are activated by the securing of likely costly programming. With distributed computing, these expenses are altogether decreased or simply vanish reliably with their infiltration. one of the advantages presented by the distributed computing model is that it moves the capital expenses recently designated to the securing of equipment and programming into operational expenses drafted by leasing the foundation and paying memberships for the usage of the product.

These expenses are regularly better controlled predictable with the business needs and flourishing of the endeavor. Distributed computing likewise presents decreases in authoritative and support costs. the amount of cost reserve funds that distributed computing can present inside an undertaking is said to the particular situation during which cloud administrations are utilized and the manner in which they add to get a benefit for the venture. inside the instance of a little startup, it's conceivable to totally use the cloud for a few angles, for example,

•IT foundation

• Software improvement

• CRM and ERP

For this situation, it's conceivable to totally dispose of capital expenses in light of the fact that there are no underlying IT resources. things are entirely unexpected inside the instance of undertakings that have just got a generous measure of IT resources. during this case, distributed computing, particularly IaaS-based arrangements, can help oversee impromptu capital costs that are created by the prerequisites of the attempt privileged the current instant. during this case, by utilizing distributed computing, these expenses are frequently turning out

to be operational costs that keep going as long as there's a prerequisite for them. for example, IT foundation renting helps all the more productively oversee top burdens without actuating capital costs. As soon in light of the fact that the expanded burden doesn't legitimize the use of extra assets, these are regularly discharged and consequently the costs identified with them vanish. this is frequently the premier embraced model of distributed computing in light of the fact that numerous undertakings have just got IT offices. an elective decision is to frame a moderate progress toward cloud-based arrangements while the capital IT resources get deteriorated and wish to get supplanted. Between these two cases, there's a decent kind of situation during which distributed computing may be of help in producing benefits for endeavors. Regarding the valuing models presented by distributed computing, we will recognize three unique methodologies that are received by the suppliers.

Reference

Cloud Computing(Concepts, Technology & Architecture) by Thomas Erl,Zaigham Mahmood, and Ricardo Puttini

(4). The pay-per-use monitor receives a "stop" event notification from the resource software (3.1) and stores the value timestamp in the log database (6).

Figure3.1.6 illustrates the pay-per-use monitor designed as a monitoring agent that transparently intercepts and analyzes runtime communication with a cloud service.

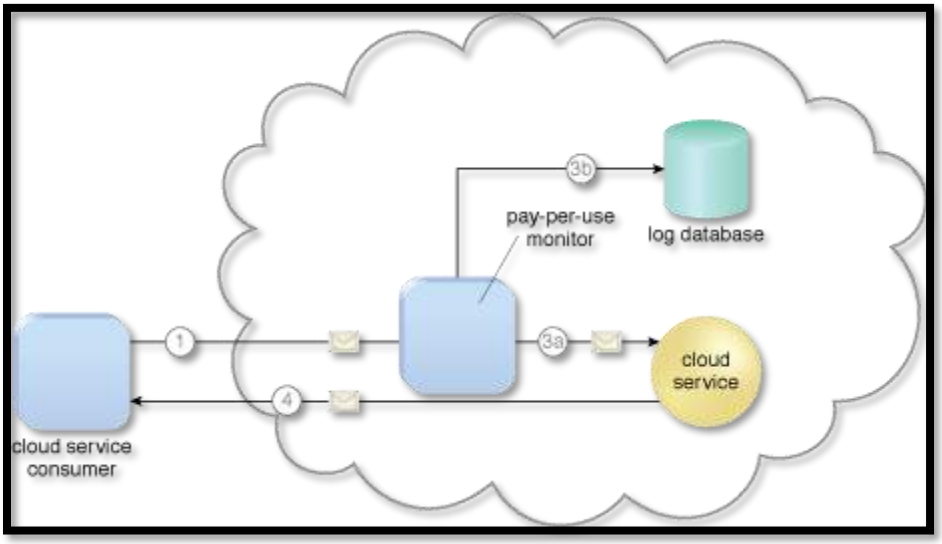


Fig 3.1.6.2

Figure 3.1.6.2 – A cloud service consumer sends a request message to the cloud service (1). The pay-per-use monitor intercepts the message (2), forwards it to the cloud service (3a), and stores the usage information in accordance with its monitoring metrics (3b). The cloud service forwards the response messages back to the cloud service consumer to provide the requested service (4).

3.1.7 Audit monitor:

The audit monitor mechanism is used to collect audit tracking data for networks and IT resources in support of, or dictated by, regulatory and contractual obligations. The figure depicts an audit monitor implemented as a monitoring agent that intercepts "login" requests and stores the requestor's security credentials, as well as both failed and successful login attempts, in a log database for future audit reporting purposes.

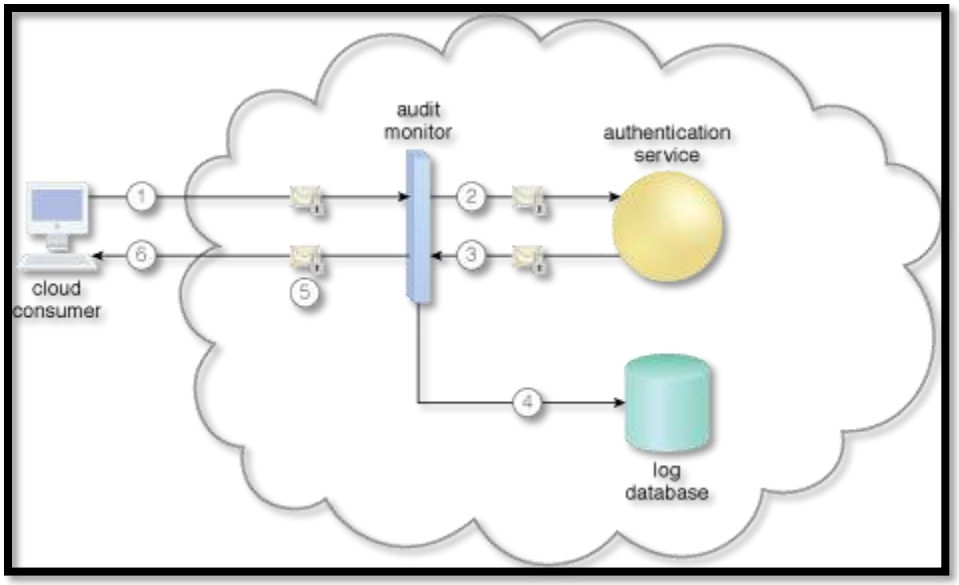


Fig 3.1.7

A cloud service consumer requests access to a cloud service by sending a login request message with security credentials (1). The audit monitor intercepts the message (2) and forwards it to the authentication service (3). The authentication service processes the security credentials. A response message is generated for the cloud service consumer, in addition to the results from the login attempt (4). The audit monitor intercepts the response message and stores the entire collected login event details in the log database, as per the organization's audit policy requirements (3.1). Access has been granted, and a response is sent back to the cloud service consumer (6).

3.1.8 Failover System:

The failover system mechanism is used to increase the reliability and availability of IT resources by using established clustering technology to provide redundant implementations. A failover system is configured to automatically switch over to a redundant or standby IT resource instance whenever the currently active IT resource becomes unavailable.

Failover systems are commonly used for mission-critical programs or for reusable services that can introduce a single point of failure for multiple applications. A failover system can span more than one geographical region so that each location hosts one or more redundant implementations of the same IT resource.

This mechanism may rely on the resource replication mechanism to supply the redundant IT resource instances, which are actively monitored for the detection of errors and unavailability conditions.

3.1.8.1 Failover systems come in two basic configurations:

A. Active-Active

In an active-active configuration, redundant implementations of the IT resource actively serve the workload synchronously (Figure 3.1.8.1). Load balancing among active instances is required. When a failure is detected, the failed instance is removed from the load balancing scheduler (Figure 3.1.8.2). Whichever IT resource remains operational when a failure is detected takes over the processing (Figure 3.1.8.3).

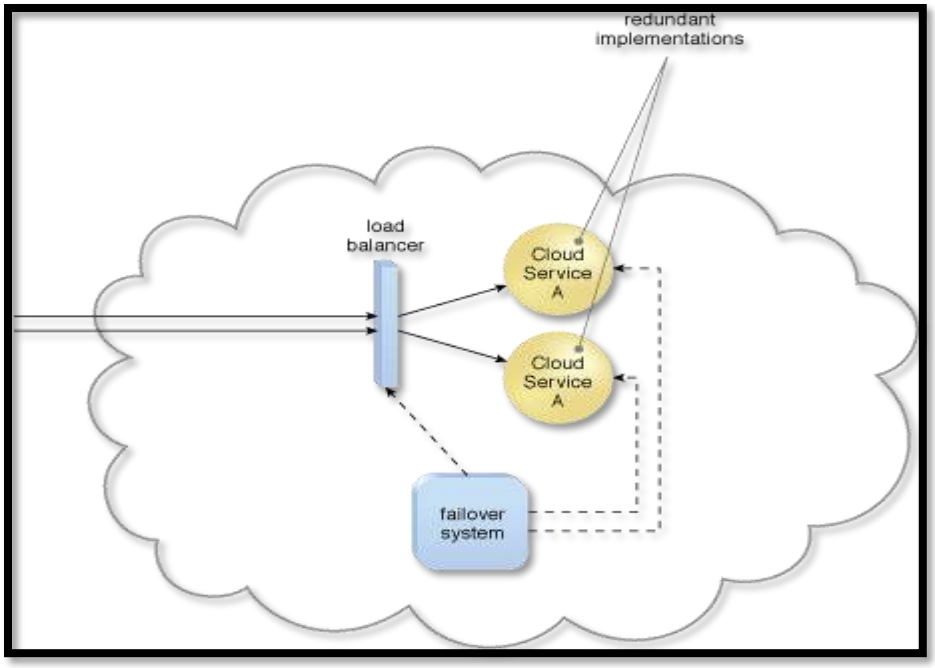


Fig 3.1.8.1

Figure 3.1.8.1 – The failover system monitors the operational status of Cloud Service A.

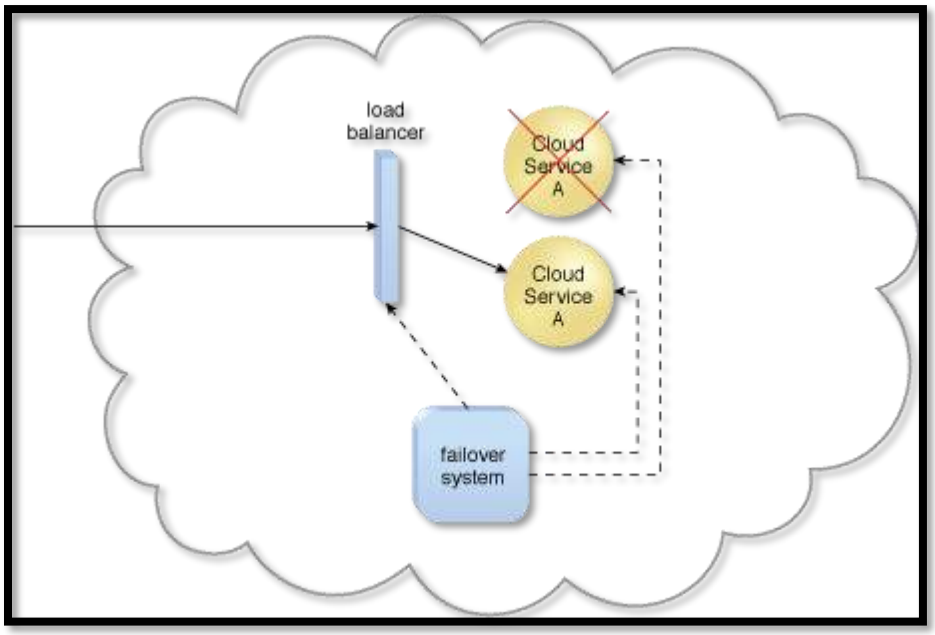


Fig 3.1.8.2

Figure 3.1.8.2 – When a failure is detected in one Cloud Service A implementation, the failover system commands the load balancer to switch over the workload to the redundant Cloud Service A implementation.

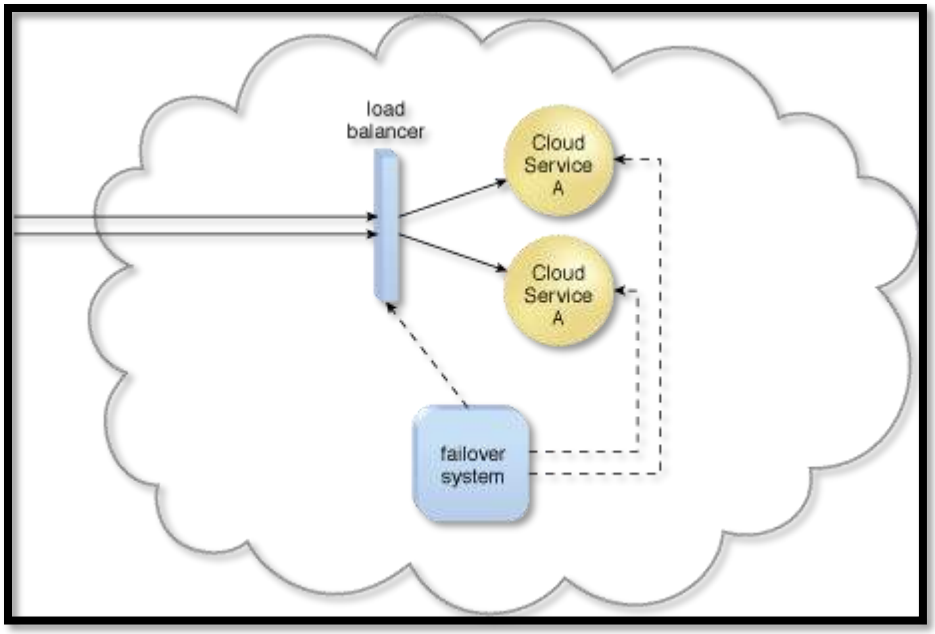


Fig 3.1.8.3

Figure 3.1.8.3 – The failed Cloud Service A implementation is recovered or replicated into an operational cloud service. The failover system now commands the load balancer to distribute the workload again.

B. Active-Passive

In an active-passive configuration, a standby or inactive implementation is activated to take over the processing from the IT resource that becomes unavailable, and the corresponding workload is redirected to the instance taking over the operation (Figures 4 to 3.1).

Some failover systems are designed to redirect workloads to active IT resources that rely on specialized load balancers that detect failure conditions and exclude failed IT resource instances from the workload distribution. This type of failover system is suitable for IT resources that do not require execution state management and provide stateless processing capabilities. In technology architectures that are typically based on clustering and virtualization technologies, the redundant or standby IT resource implementations are also required to share their state and execution context. A complex task that was executed on a failed IT resource can remain operational in one if it's redundant implementations.

While the user-interface provided by the remote administration system will tend to be proprietary to the cloud provider, there is a preference among cloud consumers to work with remote administration systems that offer standardized APIs. This allows a cloud consumer to invest in the creation of its own front-end with the foreknowledge that it can reuse this console if it decides to move to another cloud provider that supports the same standardized API. Additionally, the cloud consumer would be able to further leverage standardized APIs if it is interested in leasing and centrally administering IT resources from multiple cloud providers and/or IT resources residing in cloud and on-premise environments.

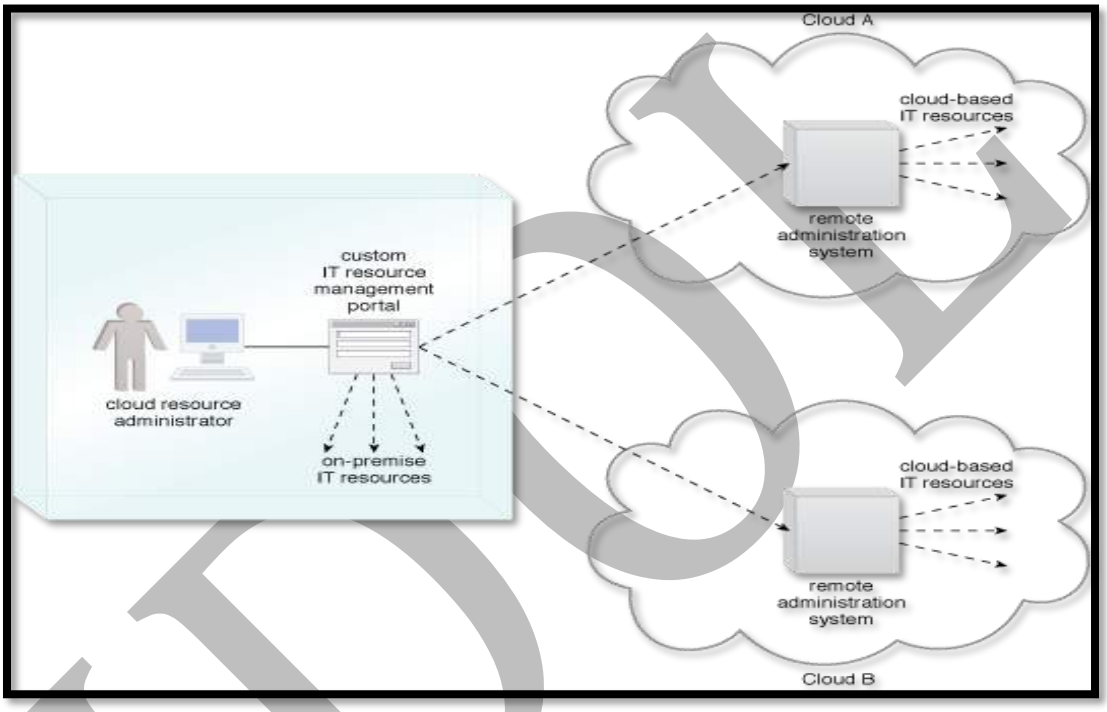


Fig 3.2.3.4

Figure 3.2.3.4 – Standardized APIs published by remote administration systems from different clouds enable a cloud consumer to develop a custom portal that centralizes a single IT resource management portal for both cloud-based and on-premise IT resources.

3.2.4 Resource Management System:

The resource management system mechanism helps coordinate IT resources in response to management actions performed by both cloud consumers and cloud providers (Figure 1). Core to this system is the virtual infrastructure manager (VIM) that coordinates the server hardware so that virtual server instances can be created from the most expedient underlying physical server. VIM is a commercial product that can be used to manage a range of virtual IT resources across multiple physical servers. For example, VIM can create and manage multiple instances of a hypervisor across different physical servers or allocate a virtual server on one physical server to another (or to a resource pool).

Tasks that are typically automated and implemented through the resource management system include:

- managing virtual IT resource templates that are used to create pre-built instances, such as virtual server images
- allocating and releasing virtual IT resources into the available physical infrastructure in response to the starting, pausing, resuming, and termination of virtual IT resource instances
- coordinating IT resources in relation to the involvement of other mechanisms, such as resource replication, load balancer, and failover system
- enforcing usage and security policies throughout the lifecycle of cloud service instances
- monitoring operational conditions of IT resources

Resource management system functions can be accessed by cloud resource administrators employed by the cloud provider or cloud consumer. Those working on behalf of a cloud provider will often be able to directly access the resource management system's native console.

Resource management systems typically expose APIs that allow cloud providers to build remote administration system portals that can be customized to selectively offer resource management controls to external cloud resource administrators acting on behalf of cloud consumer organizations via usage and administration portals.

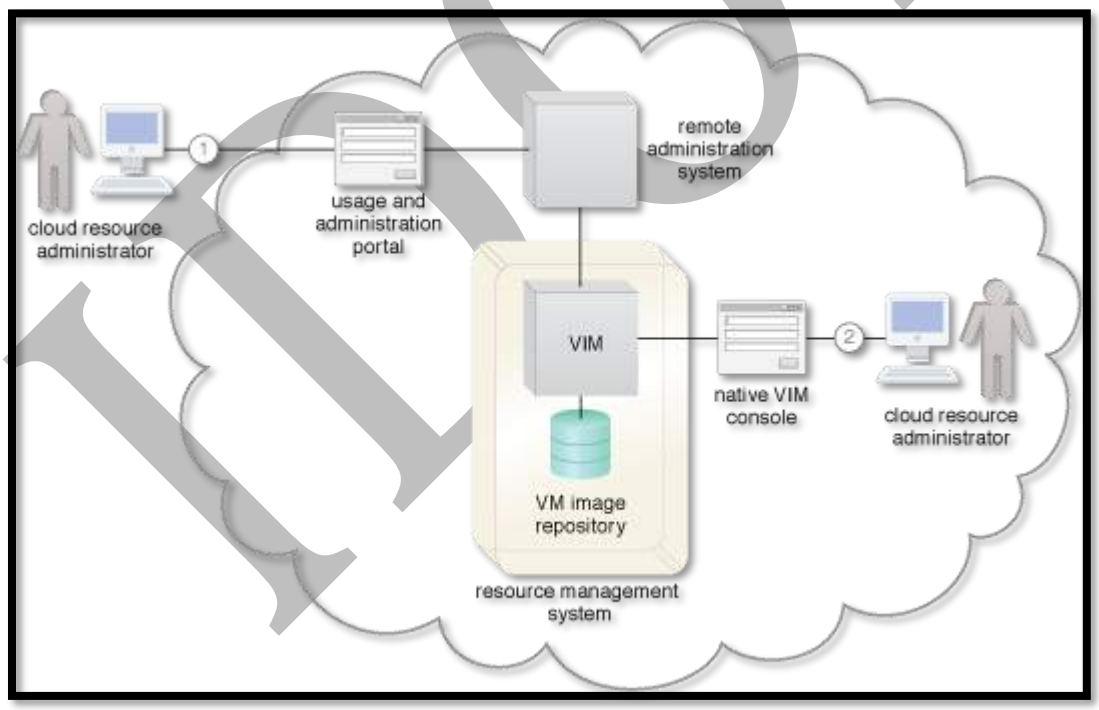


Fig 3.2.4

Figure 3.2.4 – The cloud consumer's cloud resource administrator accesses a usage and administration portal externally to administer a leased IT resource (1). The cloud provider's cloud resource administrator uses the native user-interface provided by the VIM to perform internal resource management tasks (2).

3.2.5 SLA Management System:

The SLA monitor mechanism is used to specifically observe the runtime performance of cloud services to ensure that they are fulfilling the contractual QoS requirements published in SLAs (Figure 1). The data collected by the SLA monitor is processed by an SLA management system to be aggregated into SLA reporting metrics. These systems can proactively repair or failover cloud services when exception conditions occur, such as when the SLA monitor reports a cloud service as "down."

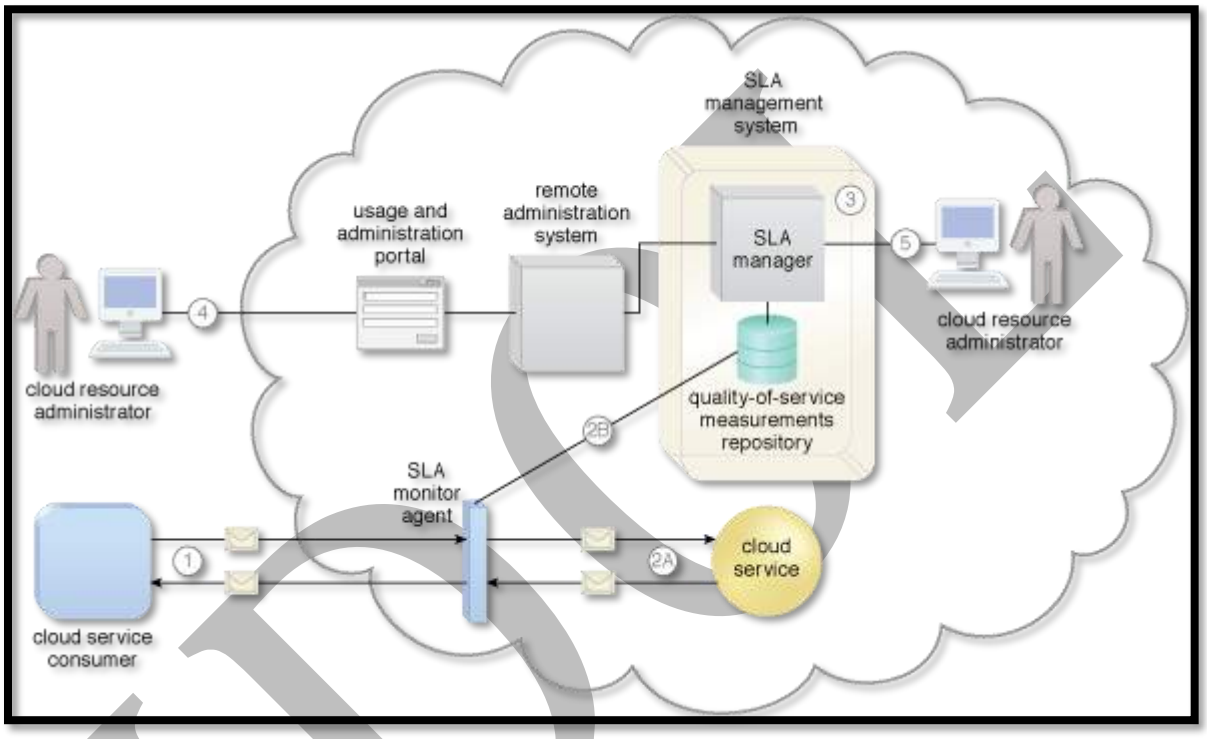


Fig 3.2.5

Figure 3.2.5 – The SLA monitor polls the cloud service by sending over polling request messages (MREQ1 to MREQN). The monitor receives polling response messages (M to M) that report that the service was "up" at each polling cycle (1a). The SLA monitor stores the "up" time—time period of all polling cycles 1 to N—in the log database (1b). The SLA monitor polls the cloud service that sends polling request messages (M to M). Polling response messages are not received (2a). The response messages continue to time out, so the SLA monitor stores the "down" time—time period of all polling cycles N+1 to N+M—in the log database (2b). The SLA monitor sends a polling request message (M) and receives the polling response message (M) (3a). The SLA monitor stores the "up" time in the log database (3b).

3.2.3.2 Billing Management System:

The billing management system mechanism is dedicated to the collection and processing of usage data as it pertains to cloud provider accounting and cloud consumer billing. Specifically, the billing management system relies on pay-per-use monitors to gather runtime usage data that is stored in a

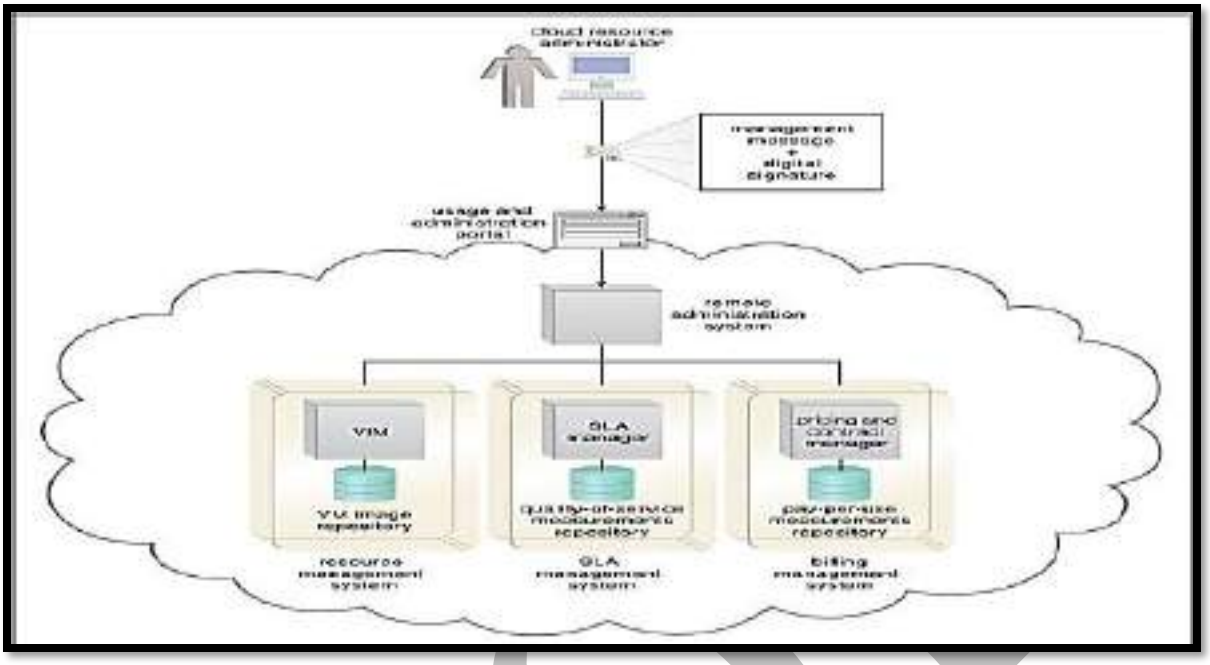


Fig 3.2.9.2

Figure 3.2.9.2 Whenever a cloud consumer performs a management action that is related to IT resources provisioned by DTGOV, the cloud service consumer program must include a digital signature in the message request to prove the legitimacy of its user.

3.2.10 Public Key Infrastructure (PKI):

- Public Key Infrastructure (PKI) A common approach for managing the issuance of asymmetric keys is based on the public key infrastructure (PKI) mechanism, which exists as a system of protocols, data formats, rules, and practices that enable large-scale systems to securely use public key cryptography.
- This system is used to associate public keys with their corresponding key owners (known as public key identification) while enabling the verification of key validity.
- PKIs rely on the use of digital certificates, which are digitally signed data structures that bind public keys to certificate owner identities, as well as to related information, such as validity periods. Digital certificates are usually digitally signed by a third-party certificate authority (CA), as illustrated in Figure 7.

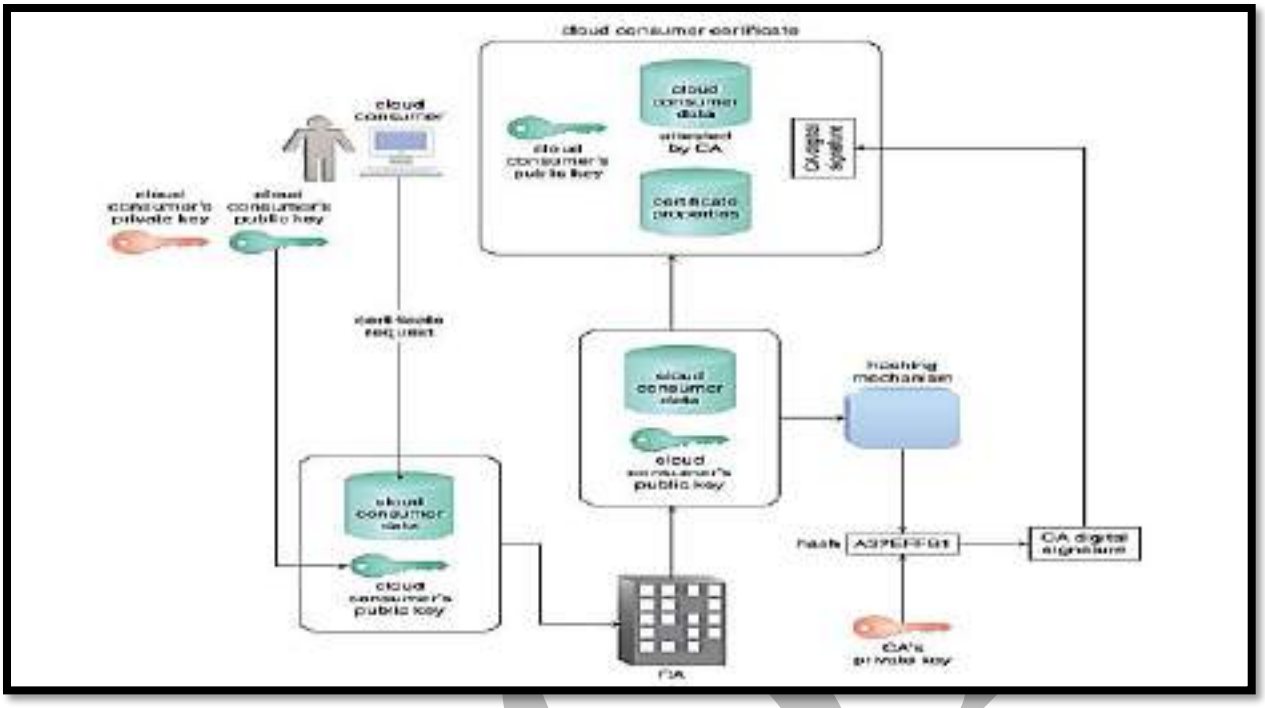


Fig 3.2.10.1

Figure 3.2.10.1 The common steps involved during the generation of certificates by a certificate authority.

- Public Key Infrastructure (PKI) Larger organizations, such as Microsoft, can act as their own CA and issue certificates to their clients and the public, since even individual users can generate certificates as long as they have the appropriate software tools.
- The PKI is a dependable method for implementing asymmetric encryption, managing cloud consumer and cloud provider identity information, and helping to defend against the malicious intermediary and insufficient authorization threats.
- The PKI mechanism is primarily used to counter the insufficient authorization threat.

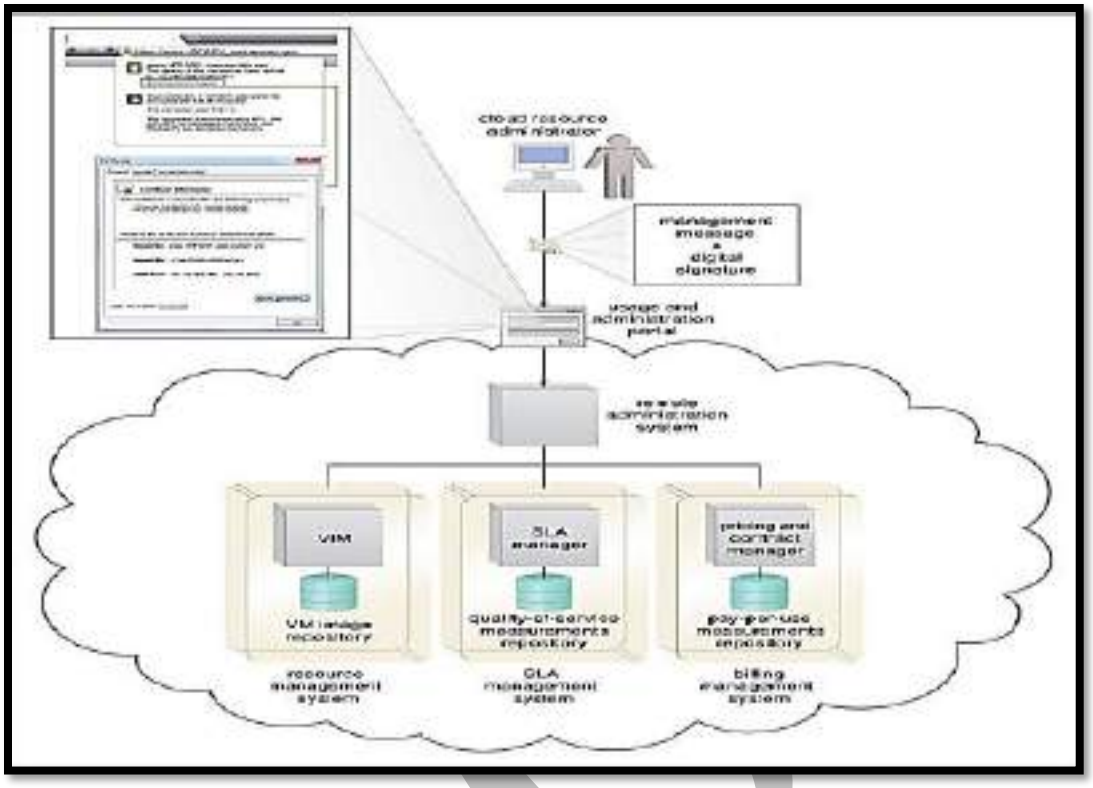


Fig 3.2.10.2

Figure 3.2.10.2 An external cloud resource administrator uses a digital certificate to access the Web-based management environment. DTGOV's digital certificate is used in the HTTPS connection and then signed by a trusted CA.

3.2.11 Identity and Access Management (IAM):

Identity and Access Management (IAM) The Identity and access management (IAM) mechanism encompasses the components and policies necessary to control and track user identities and access privileges for IT resources, environments, and systems. Specifically, IAM mechanisms exist as systems comprised of four main components:

1. **Authentication –** Username and password combinations remain the most common forms of user authentication credentials managed by the IAM system, which also can support digital signatures, digital certificates, biometric hardware (fingerprint readers), specialized software (such as voice analysis programs), and locking user accounts to registered IP or MAC addresses.
2. **Authorization –** The authorization component defines the correct granularity for access controls and oversees the relationships between identities, access control rights, and IT resource availability.
3. **User Management –** Related to the administrative capabilities of the system, the user management program is responsible for creating new user identities and access groups, resetting passwords, defining password policies, and managing privileges.

4. **Credential Management –** The credential management system establishes identities and access control rules for defined user accounts, which mitigates the threat of insufficient authorization. The IAM mechanism is primarily used to counter the insufficient authorization, denial of service, and overlapping trust boundaries threats.

3.2.12 Single Sign-On (SSO):

- Single Sign-On (SSO) Propagating the authentication and authorization information for a cloud service consumer across multiple cloud services can be a challenge, especially if numerous cloud services or cloud-based IT resources need to be invoked as part of the same overall runtime activity.
- The single sign-on (SSO) mechanism enables one cloud service consumer to be authenticated by a security broker, which establishes a security context that is persisted while the cloud service consumer accesses other cloud services or cloud-based IT resources.
- Otherwise, the cloud service consumer would need to re-authenticate itself with every subsequent request. The SSO mechanism essentially enables mutually independent cloud services and IT resources to generate and circulate runtime authentication and authorization credentials.

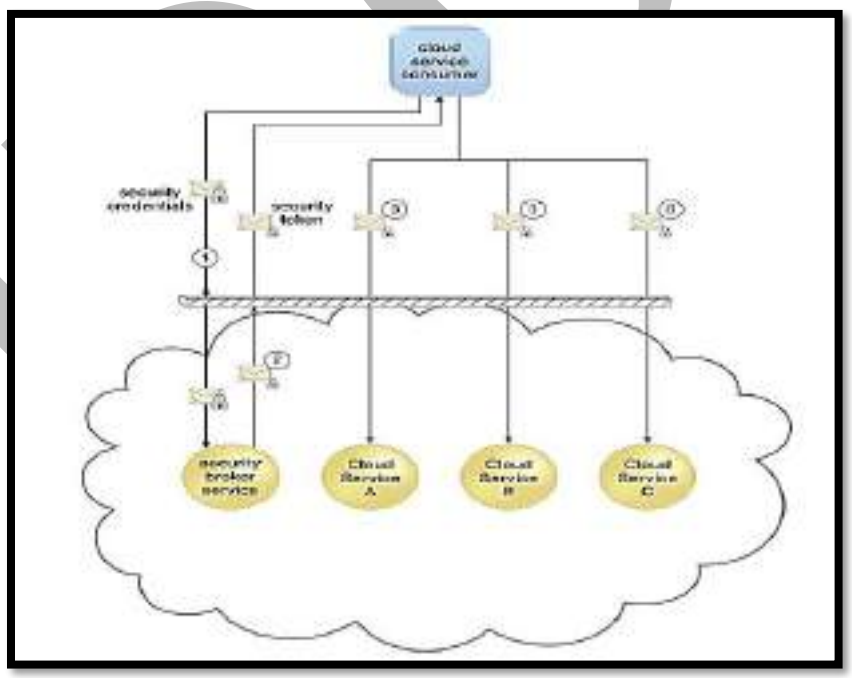
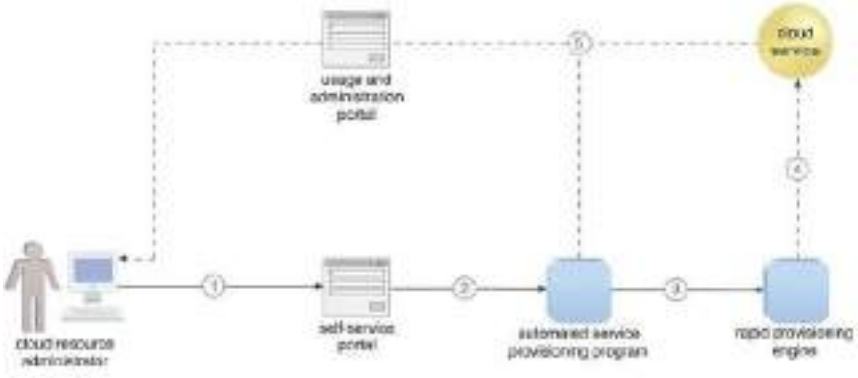


Fig 3.2.12.1

Figure 3.2.12.1 A cloud service consumer provides the security broker with login credentials (1). The security broker responds with an authentication token (message with small lock symbol) upon successful authentication, which contains cloud service consumer identity information (2) that is

8.9. Rapid Provisioning Architecture

The *rapid provisioning architecture* establishes a system that automates the provisioning of a wide range of resources, either individually or as a collective. The underlying technology architecture for rapid resource provisioning can be sophisticated and complex, and relies on a system comprised of an automated provisioning program, rapid provisioning engine, and scripts and templates for on-demand provisioning.



(1) A cloud resource administrator requests a new cloud service through the self-service portal.
(2) The self-service portal passes the request to the automated service provisioning program installed on the virtual server.
(3) which passes the necessary tasks to be performed to the rapid provisioning engine.
(4) The rapid provisioning engine announces when the new cloud service is ready.
(5) The automated service provisioning program finalizes and publishes the cloud service on the usage and administration portal for cloud consumer access.

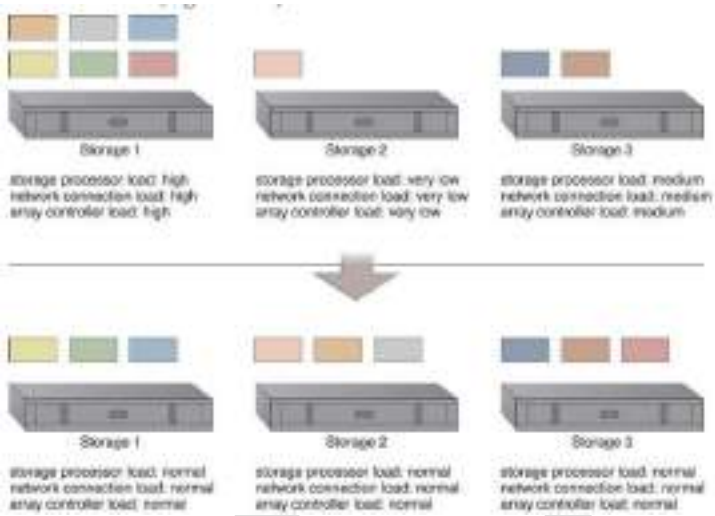
The step-by-step description describes the inner workings of a rapid provisioning engine:

1. A cloud consumer requests a new server through the self-service portal.
2. The sequence manager forwards the request to the deployment engine for the preparation of an operating system.
3. The deployment engine uses the virtual server templates for provisioning if the request is for a virtual server. Otherwise, the deployment engine sends the request to provision a physical server.
4. The pre-defined image for the requested type of operating system is used for the provisioning of the operating system, if available. Otherwise, the regular deployment process is executed to install the operating system.
5. The deployment engine informs the sequence manager when the operating system is ready.
6. The sequence manager updates and sends the logs to the sequence logger for storage.
7. The sequence manager requests that the deployment engine apply the operating system baseline to the provisioned operating system.
8. The deployment engine applies the requested operating system baseline.
9. The deployment engine informs the sequence manager that the operating system baseline has been applied.
10. The sequence manager updates and sends the logs of completed steps to the sequence logger for storage.
11. The sequence manager requests that the deployment engine install the applications.
12. The deployment engine deploys the applications on the provisioned server.
13. The deployment engine informs the sequence manager that the applications have been installed.
14. The sequence manager updates and sends the logs of completed steps to the sequence logger for storage.
15. The sequence manager requests that the deployment engine apply the application's configuration baseline.

16. The deployment engine applies the configuration baseline.
17. The deployment engine informs the sequence manager that the configuration baseline has been applied.
18. The sequence manager updates and sends the logs of completed steps to the sequence logger for storage.

8.10. Storage Workload Management Architecture

This architecture enables LUNs to be evenly distributed across available cloud storage devices, while a storage capacity system is established to ensure that runtime workloads are evenly distributed across the LUNs.



Combining cloud storage devices into a group allows LUN data to be distributed between available storage hosts equally. A storage management system is configured and an automated scaling listener is positioned to monitor and equalize runtime workloads among the grouped cloud storage devices.

Glossary:

- **Intelligent Automation Engine:** The intelligent automation engine automates administration tasks by executing scripts that contain workflow logic.
- **LUN:** A logical unit number (LUN) is a logical drive that represents a partition of a physical drive.
- **Storage Service Gateway:** The storage service gateway is a component that acts as the external interface to cloud storage services, and is capable of automatically redirecting cloud consumer requests whenever the location of the requested data has changed.
- **Storage Replication:** Storage replication is a variation of the resource replication mechanisms used to synchronously or asynchronously replicate data from a primary storage device to a secondary storage device. It can be used to replicate partial and entire LUNs.
- **Heartbeats:** Heartbeats are system-level messages exchanged between hypervisors, hypervisors and virtual servers, and hypervisors and VIMs.
- **Live VM migration:** Live VM migration is a system that is capable of relocating virtual servers or virtual server instances at runtime.
- **LUN migration:** LUN migration is a specialized storage program that is used to move LUNs from one storage device to another without interruption, while remaining transparent to cloud consumers.

References:

Cloud Computing Concepts, Technology & Architecture
- Thomas Erl, Zaigham Mahmood, and Ricardo Puttini – Prentice Hall - 2013
Chapter 11: Fundamental Cloud Architectures
Chapter 12: Advanced Cloud Architectures

Mastering Cloud Computing Foundations and Applications Programming
- Rajkumar Buyya, Christian Vecchiola, S. Thamarai Selvi - Elsevier – 2013

Distributed and Cloud Computing, From Parallel Processing to the Internet of Things
- Kai Hwang, Jack Dongarra, Geoffrey Fox – MK Publishers - 2012

Unit 4
Chapter – 1
Fundamental Cloud Architectures

Unit Structure:

4.1.1 Workload Distribution Architecture
4.2.1 Resource Pooling Architecture
4.3.1 Dynamic Scalability Architecture
4.4.1 Elastic Resource Capacity Architecture
4.5.1 Service Load Balancing Architecture
4.6.1 Cloud Bursting Architecture
4.7.1 Elastic Disk Provisioning Architecture
4.8.1 Redundant Storage Architecture

Objective:

To learn how to use Cloud Services.

Introduction:

This chapter introduces and describes several of the more common foundational cloud architectural models, each explaining a common usage and characteristic of modern day cloud-based environments. Further the chapter also explores the involvement and importance of different combinations of cloud computing mechanisms in relation to these architectures.

4.1. Workload Distribution Architecture

- Resources on cloud can be horizontally scald using an addition or identical resource and a load balancer that is capable of providing run time distribution of workload among resources.
- This architecture of distribution has a dual advantage
  - **i.** Reduces overutilization of resources.
  - **ii.** Reduces underutilization of resources.

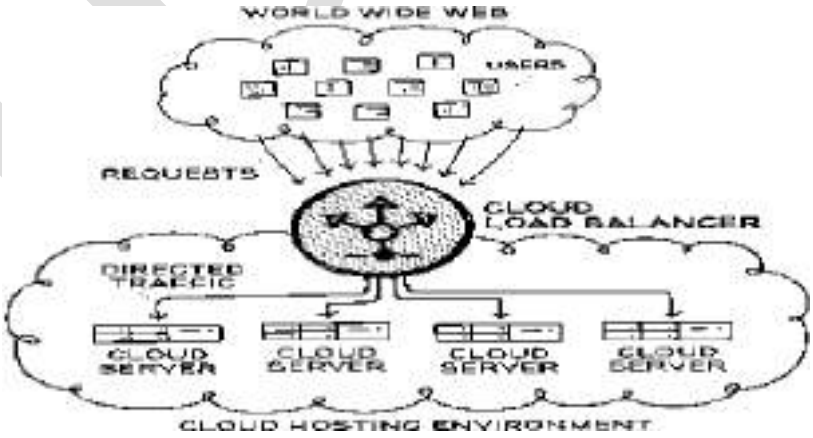
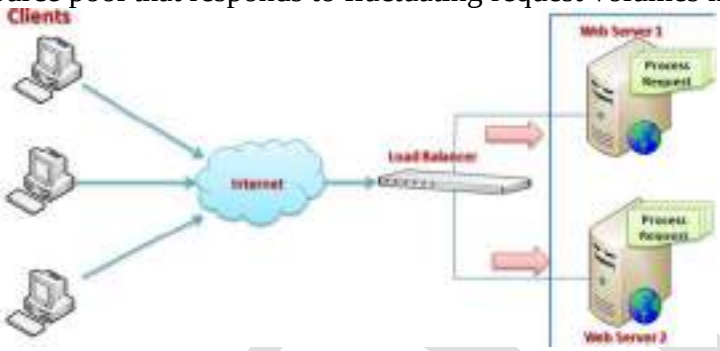


Figure: Workload Distribution Architecture

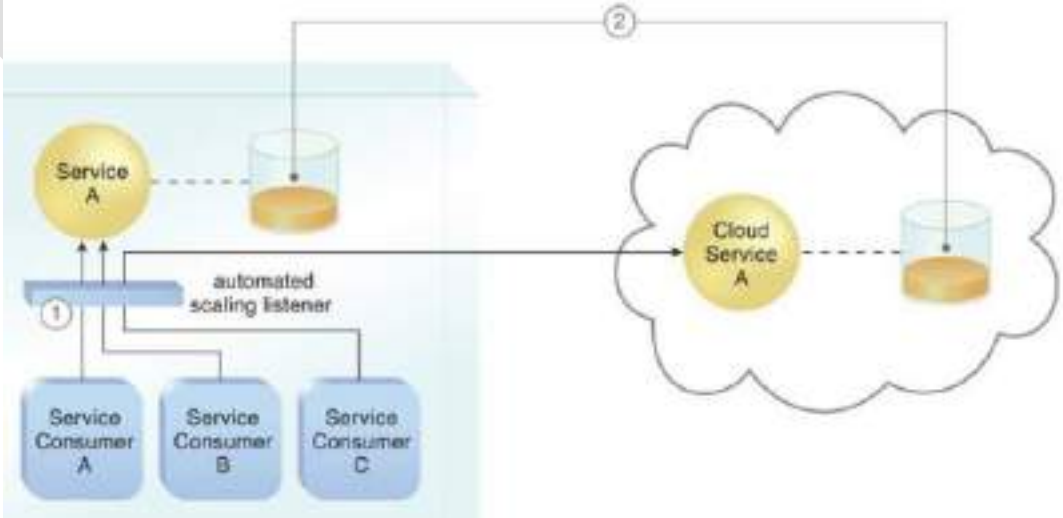
- Workload distribution is carried out in support of distributed virtual servers, storage devices and services.
- Load balancing system produces specialized variation that incorporates the aspect of load balancing like:
  - **i.** Load Balanced Service Instances Architecture

- This architecture is a specialized variant of the workload distribution architecture that is geared specifically for scaling cloud service implementations.
- Redundant deployments of cloud services are created, with a load balancing system added to dynamically distribute workloads.
- The duplicate cloud service implementations are organized into a resource pool, while the load balancer is positioned as either an external or built-in component to allow the host servers to balance the workloads themselves.
- Depending on the anticipated workload and processing capacity of host server environments, multiple instances of each cloud service implementation can be generated as part of a resource pool that responds to fluctuating request volumes more efficiently.



4.6. Cloud Bursting Architecture

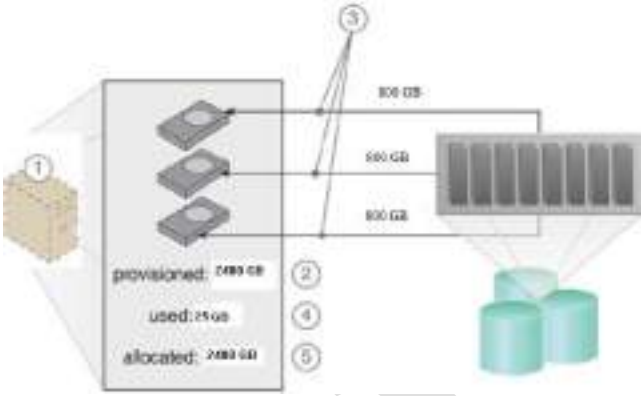
- This architecture establishes a form of dynamic scaling that scales or "bursts out" on-premise cloud resources into a cloud whenever predefined capacity thresholds have been reached.
- The corresponding cloud-based resources are redundantly pre-deployed but remain inactive until cloud bursting occurs. After they are no longer required, the resources are released and the architecture "bursts in" back to the on-premise environment.
- Cloud bursting is a flexible scaling architecture that provides cloud consumers with the option of using cloud-based IT resources only to meet higher usage demands.
- The foundation of this architectural model is based on the automated scaling listener and resource replication mechanisms.
- The automated scaling listener determines when to redirect requests to cloud based resources, and resource replication is used to maintain synchronicity between on-premise and cloud-based IT resources in relation to state information



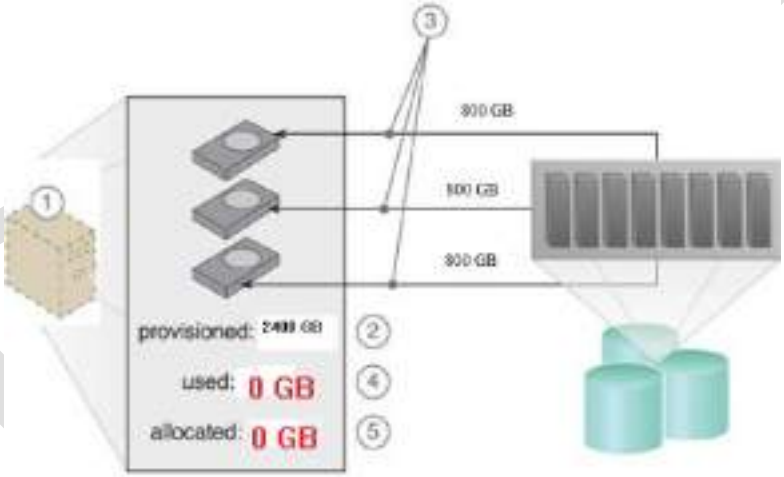
4.7. Elastic Disk Provisioning Architecture

- Cloud consumers are commonly charged for cloud-based storage space based on fixed-disk storage allocation, meaning the charges are predetermined by disk capacity and not aligned with actual data storage consumption.

- Cloud provisions a virtual server with the Windows Server 2019 OS and Three 800 GB hard drives. The cloud consumer is billed for using 2400 GB of storage space after installing the operating system, even though the operating system only requires 25 GB of storage space.



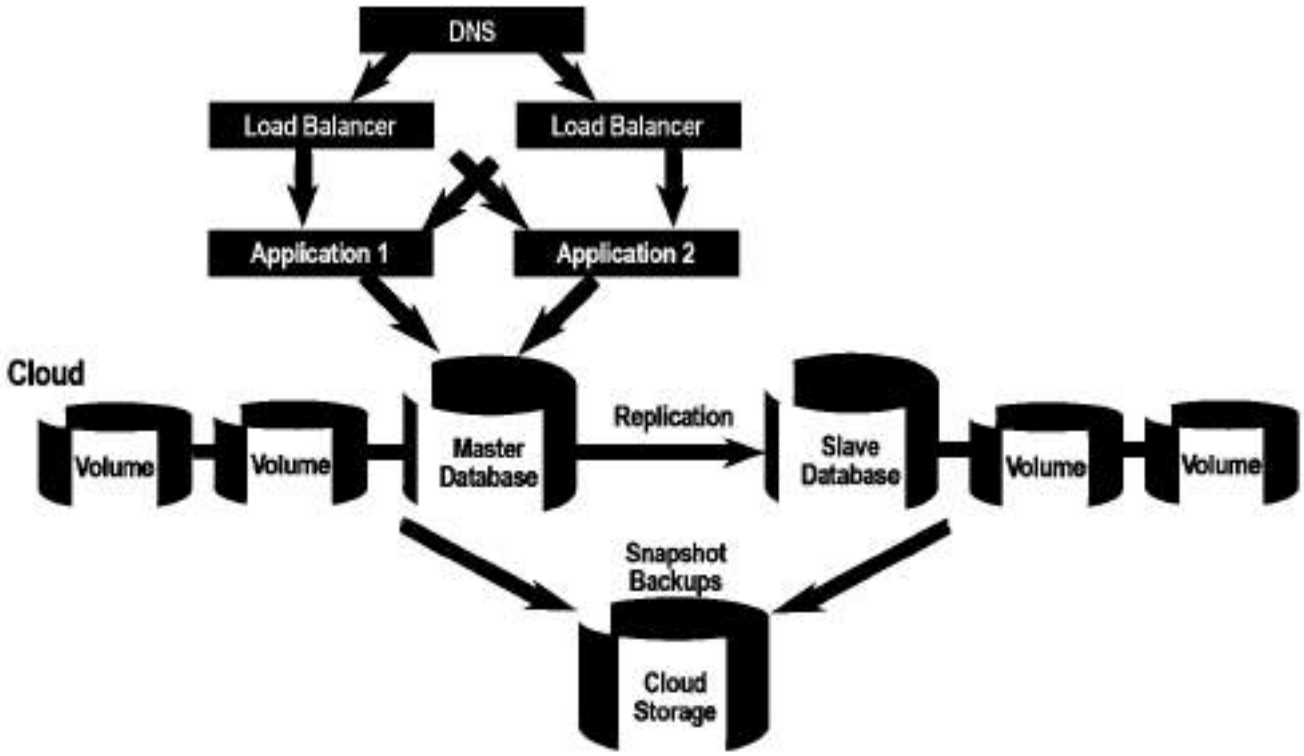
- The *elastic disk provisioning architecture* establishes a dynamic storage provisioning system that ensures that the cloud consumer is granularly billed for the exact amount of storage that it actually uses. This system uses thin provisioning technology for the dynamic allocation of storage space, and is further supported by runtime usage monitoring to collect accurate usage data for billing purposes



- Thin-provisioning software is installed on virtual servers that process dynamic storage allocation via the hypervisor, while the pay-per-use monitor tracks and reports granular billing-related disk usage data.

4.8. Redundant Storage Architecture

- Cloud storage devices are occasionally subject to failure and disruptions that are caused by network connectivity issues, controller or general hardware failure, or security breaches.
- A compromised cloud storage device's reliability can have a ripple effect and cause impact failure across all of the services, applications, and infrastructure components in the cloud that are reliant on its availability.
- The *redundant storage architecture* introduces a secondary duplicate cloud storage device as part of a failover system that synchronizes its data with the data in the primary cloud storage device.
- A storage service gateway diverts cloud consumer requests to the secondary device whenever the primary device fails.



- This cloud architecture primarily relies on a storage replication system that keeps the primary cloud storage device synchronized with its duplicate secondary cloud storage devices.
- Cloud providers may locate secondary cloud storage devices in a different geographical region than the primary cloud storage device, usually for economic reasons.
- The location of the secondary cloud storage devices can dictate the protocol and method used for synchronization, like some replication transport protocols have distance restrictions.

Chapter – 2
Advanced Cloud Architectures

Unit Structure:

4.2.1. Hypervisor Clustering Architecture
4.2.2. Load Balanced Virtual Server Instances Architecture
4.2.3. Non-Disruptive Service Relocation Architecture
4.2.4. Zero Downtime Architecture
4.2.5. Cloud Balancing Architecture
4.2.6. Resource Reservation Architecture
4.2.7. Dynamic Failure Detection and Recovery Architecture
4.2.8. Bare-Metal Provisioning Architecture
4.2.9. Rapid Provisioning Architecture
4.2.10. Storage Workload Management Architecture

Objective:

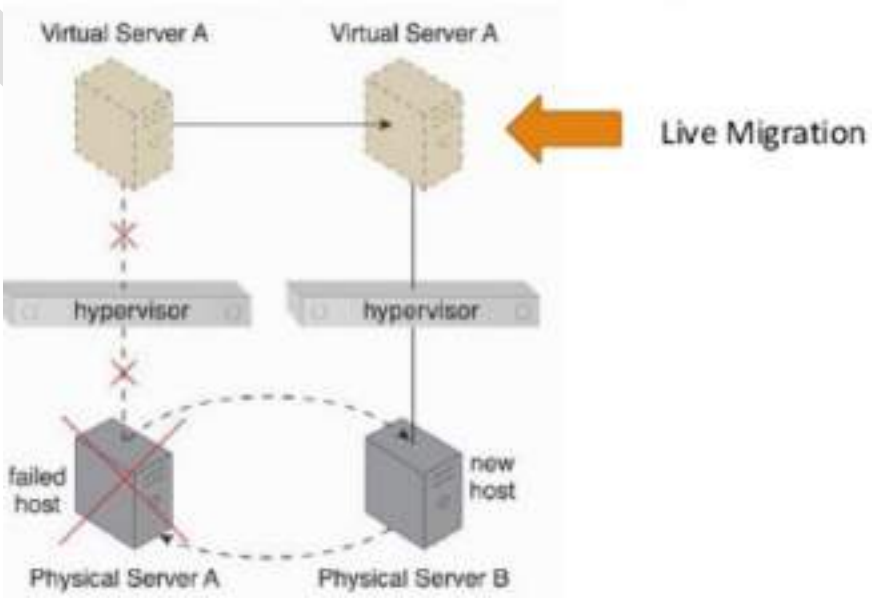
To learn how to use Advance Cloud Services.

Introduction:

This chapter introduces the cloud technology architectures distinct and sophisticated architectural layers, several of which can be built upon the more foundational environments established by the architectural models discussed in previous chapter.

4.2.1 Hypervisor Clustering Architecture

- Hypervisors are responsible for creating and hosting multiple virtual servers.
- Because of this dependency, any failure conditions that affect a hypervisor can cascaded effect on its virtual servers.
- The *hypervisor clustering architecture* establishes a high-availability cluster of hypervisors across multiple physical servers.
- If a given hypervisor or its underlying physical server becomes unavailable, the hosted virtual servers can be moved to another physical server or hypervisor to maintain runtime operations.



- The hypervisor cluster is controlled via a central VIM, which sends regular heartbeat messages to the hypervisors to confirm that they are up and running.

Glossary:

- **Intelligent Automation Engine:** The intelligent automation engine automates administration tasks by executing scripts that contain workflow logic.
- **LUN:** A logical unit number (LUN) is a logical drive that represents a partition of a physical drive.
- **Storage Service Gateway:** The storage service gateway is a component that acts as the external interface to cloud storage services, and is capable of automatically redirecting cloud consumer requests whenever the location of the requested data has changed.
- **Storage Replication:** Storage replication is a variation of the resource replication mechanisms used to synchronously or asynchronously replicate data from a primary storage device to a secondary storage device. It can be used to replicate partial and entire LUNs.
- **Heartbeats:** Heartbeats are system-level messages exchanged between hypervisors, hypervisors and virtual servers, and hypervisors and VIMs.
- **Live VM migration:** Live VM migration is a system that is capable of relocating virtual servers or virtual server instances at runtime.
- **LUN migration:** LUN migration is a specialized storage program that is used to move LUNs from one storage device to another without interruption, while remaining transparent to cloud consumers.

References:

Cloud Computing Concepts, Technology & Architecture
- Thomas Erl, Zaigham Mahmood, and Ricardo Puttini – Prentice Hall - 2013
Chapter 11: Fundamental Cloud Architectures
Chapter 12: Advanced Cloud Architectures

Mastering Cloud Computing Foundations and Applications Programming
- Rajkumar Buyya, Christian Vecchiola, S. Thamarai Selvi - Elsevier – 2013

Distributed and Cloud Computing, From Parallel Processing to the Internet of Things
- Kai Hwang, Jack Dongarra, Geoffrey Fox – MK Publishers - 2012

Unit 5: Chapter 1
Cloud Delivery Model Considerations

Unit Structure

5.0 OBJECTIVES

- Describe cloud delivery models for IaaS
- Describe cloud delivery models for PaaS
- Describe cloud delivery models for SaaS
- Describe different ways in which cloud delivery models are administered and utilized by cloud consumers
- Working with IaaS Environments
- Working with PaaS Environments
- Working with SaaS Environments

5.1 INTRODUCTION

A cloud delivery model represents a specific combination of IT resources offered by a cloud provider. This terminology is typically associated with cloud computing and frequently used to describe a type of remote environment and the level of control.

5.2 Cloud Delivery Models: The Cloud Provider Perspective

This section explores the architecture and administration of IaaS, PaaS, and SaaS cloud delivery models from the point of view of the cloud provider (Figure 5.1). The integration and management of these cloud-based environments as part of greater environments and how they can relate to different technologies and cloud mechanism combinations are examined.

Figure 5.1

5.2.1 Building IaaS Environments

The virtual server and cloud storage device mechanisms represent the two most fundamental IT resources that are delivered as part of a standard rapid provisioning architecture within IaaS environments. They are offered in various standardized configurations that are defined by the following properties:

- Operating System
- Primary Memory Capacity
- Processing Capacity
- Virtualized Storage Capacity

Memory and virtualized storage capacity is usually allocated with increments of 1 GB to simplify the provisioning of underlying physical IT resources. When limiting cloud consumer access to virtualized environments, IaaS offerings are preemptively assembled by cloud providers via virtual server images that capture the pre-defined configurations. Some cloud providers may offer cloud consumers direct administrative access to physical IT resources, in which case the bare-metal provisioning architecture may come into play.

Snapshots can be taken of a virtual server to record its current state, memory, and configuration of a virtualized IaaS environment for backup and replication purposes, in support of horizontal and vertical scaling requirements. For example, a virtual server can use its snapshot to become reinitialized in another hosting environment after its capacity has been increased to allow for vertical scaling. The snapshot can alternatively be used to duplicate a virtual server. The management of custom virtual server images is a vital feature that is provided via the remote administration system mechanism. Most cloud providers also support importing and exporting options for custom-built virtual server images in both proprietary and standard formats.

Data Centers

Cloud providers can offer IaaS-based IT resources from multiple geographically diverse data centers, which provides the following primary benefits:

- Multiple data centers can be linked together for increased resiliency. Each data center is placed in a different location to lower the chances of a single failure forcing all of the data centers to go offline simultaneously.
- Connected through high-speed communications networks with low latency, data centers can perform load balancing, IT resource backup and replication, and increase storage capacity, while improving availability and reliability. Having multiple data centers spread over a greater area further reduces network latency.
- Data centers that are deployed in different countries make access to IT resources more convenient for cloud consumers that are constricted by legal and regulatory requirements.

When an IaaS environment is used to provide cloud consumers with virtualized network environments, each cloud consumer is segregated into a tenant environment that isolates IT resources from the rest of the cloud through the Internet. VLANs and network access control software collaboratively realize the corresponding logical network perimeters.

Scalability and Reliability

Within IaaS environments, cloud providers can automatically provision virtual servers via the dynamic vertical scaling type of the dynamic scalability architecture. This can be performed through the VIM, as long as the host physical servers have sufficient capacity. The VIM can scale virtual servers out using resource replication as part of a resource pool architecture, if a given physical server has insufficient capacity to support vertical scaling. The load balancer mechanism, as part of a workload distribution architecture, can be used to distribute the workload among IT resources in a pool to complete the horizontal scaling process.

Manual scalability requires the cloud consumer to interact with a usage and administration program to explicitly request IT resource scaling. In contrast, automatic scalability requires the automated scaling listener to monitor the workload and reactively scale the resource capacity. This mechanism typically acts as a monitoring agent that tracks IT resource usage in order to notify the resource management system when capacity has been exceeded.

Replicated IT resources can be arranged in high-availability configuration that forms a failover system for implementation via standard VIM features. Alternatively, a high-availability/high-performance resource cluster can be created at the physical or virtual server level, or both simultaneously. The multipath resource access architecture is commonly employed to enhance reliability via the use of redundant access paths, and some cloud providers further offer the provisioning of dedicated IT resources via the resource reservation architecture.

Monitoring

Cloud usage monitors in an IaaS environment can be implemented using the VIM or specialized monitoring tools that directly comprise and/or interface with the virtualization platform. Several common capabilities of the IaaS platform involve monitoring:

- ***Virtual Server Lifecycles*** - Recording and tracking uptime periods and the allocation of IT resources, for pay-per-use monitors and time-based billing purposes.
- ***Data Storage*** - Tracking and assigning the allocation of storage capacity to cloud storage devices on virtual servers, for pay-per-use monitors that record storage usage

backend architecture for usage measurement and billing- related data collection is determined and implemented, including the positioning of pay-per-use monitor and billing management system mechanisms.

- ***Cloud Service Contracting*** - This phase consists of negotiations between the cloud consumer and cloud provider with the goal of reaching a mutual agreement on rates based on usage cost metrics.
- ***Cloud Service Offering*** - This stage entails the concrete offering of a cloud service's pricing models through cost templates, and any available customization options.
- ***Cloud Service Provisioning*** - Cloud service usage and instance creation thresholds may be imposed by the cloud provider or set by the cloud consumer. Either way, these and other provisioning options can impact usage costs and other fees.
- ***Cloud Service Operation*** - This is the phase during which active usage of the cloud service produces usage cost metric data.
- ***Cloud Service Decommissioning*** - When a cloud service is temporarily or permanently deactivated, statistical cost data may be archived.

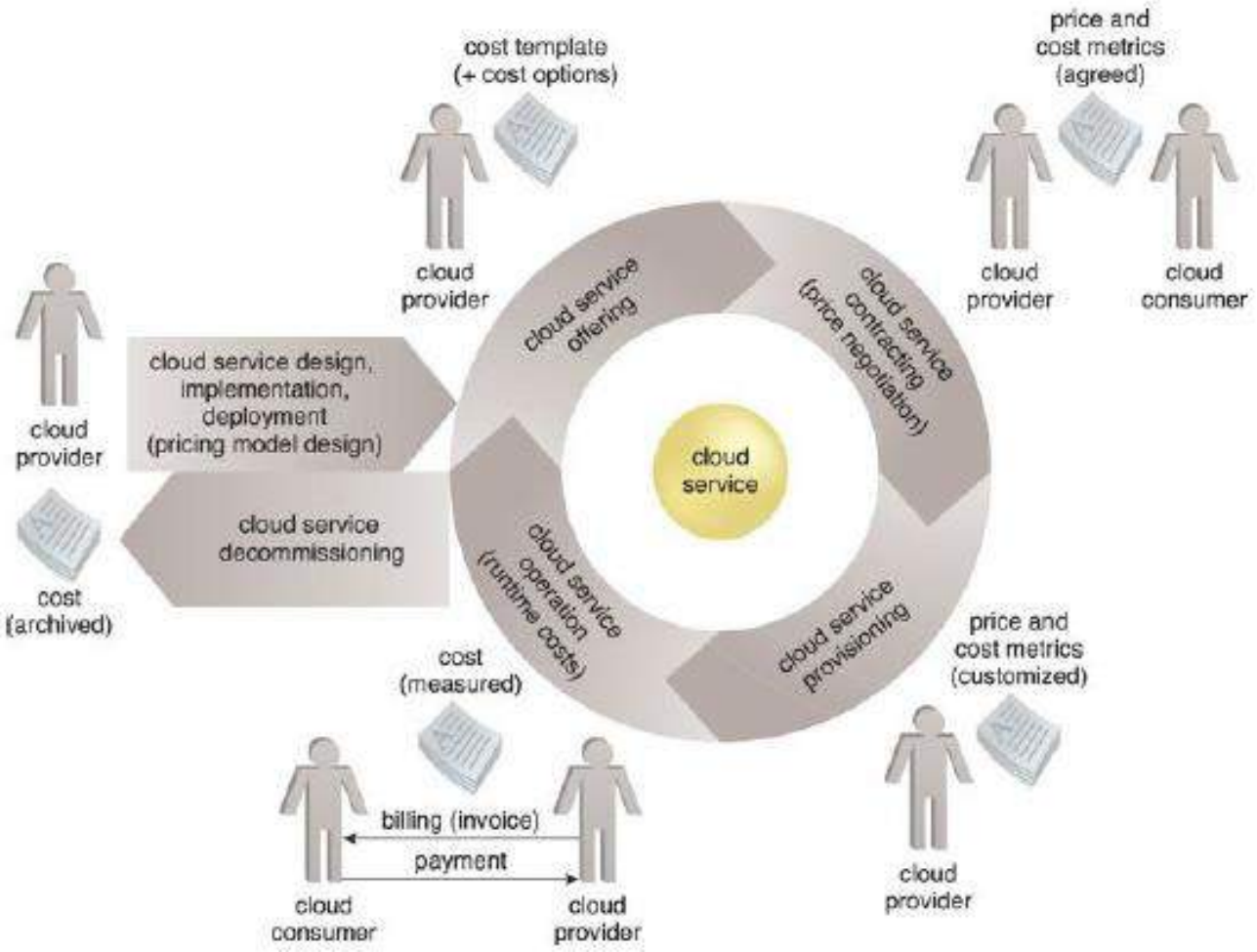


Figure 5.2..1 Common cloud service lifecycle stages as they relate to cost management considerations.

5.2..4.1 Pricing Models

5.2..6.1 Service Availability Metrics

Availability Rate Metric

The overall availability of an IT resource is usually expressed as a percentage of up-time. For example, an IT resource that is always available will have an uptime of 5.2.0%.

- ***Description*** - percentage of service up-time
- ***Measurement*** - total up-time / total time
- ***Frequency*** - weekly, monthly, yearly
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - minimum 55.5% up-time

Availability rates are calculated cumulatively, meaning that unavailability periods are combined in order to compute the total downtime (Table 5.2..1)

Availability (%)	Downtime/Week (Seconds)	Downtime/Month (Seconds)	Downtime/Year (Seconds)
99.5	3024	216	158112
99.8	1210	5174	63072
99.9	606	2592	31536
99.95	302	1294	15768
99.99	60.6	259.2	3154
99.999	6.05	25.9	316.6
99.9999	0.605	2.59	31.5

Table 5.2..1 Sample availability rates measured in units of seconds

Outage Duration Metric

This service quality metric is used to define both maximum and average continuous outage service-level targets.

- ***Description*** - duration of a single outage
- ***Measurement*** - date/time of outage end - date/time of outage start
- ***Frequency*** - per event
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 1 hour maximum, 15 minutes average

5.2..6.2 Service Reliability Metrics

A characteristic closely related to availability; reliability is the probability that an IT resource can perform its intended function under pre-defined conditions without experiencing failure. Reliability focuses on how often the service performs as expected, which requires the service to remain in an operational and available state. Certain reliability metrics only consider runtime

errors and exception conditions as failures, which are commonly measured only when the IT resource is available.

Mean-Time Between Failures (MTBF) Metric

- ***Description*** - expected time between consecutive service failures
- ***Measurement*** - £, normal operational period duration / number of failures
- ***Frequency*** - monthly, yearly
- ***Cloud Delivery Model*** - IaaS, PaaS
- ***Example*** - 50 day average

Reliability Rate Metric

Overall reliability is more complicated to measure and is usually defined by a reliability rate that represents the percentage of successful service outcomes.

This metric measures the effects of non-fatal errors and failures that occur during up-time periods. For example, an IT resource's reliability is 5.2.0% if it has performed as expected every time it is invoked, but only 80% if it fails to perform every fifth time.

- ***Description*** - percentage of successful service outcomes under pre-defined conditions
- ***Measurement*** - total number of successful responses / total number of requests
- ***Frequency*** - weekly, monthly, yearly
- ***Cloud Delivery Model*** - SaaS
- ***Example*** - minimum 55.5%

5.2..6.3 Service Performance Metrics

Service performance refers to the ability on an IT resource to carry out its functions within expected parameters. This quality is measured using service capacity metrics, each of which focuses on a related measurable characteristic of IT resource capacity. A set of common performance capacity metrics is provided in this section. Note that different metrics may apply, depending on the type of IT resource being measured.

Network Capacity Metric

- ***Description*** - measurable characteristics of network capacity
- ***Measurement*** - bandwidth / throughput in bits per second
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 5.2. MB per second

Storage Device Capacity Metric

- ***Description*** - measurable characteristics of storage device capacity
- ***Measurement*** - storage size in GB
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 80 GB of storage

Server Capacity Metric

- ***Description*** - measurable characteristics of server capacity
- ***Measurement*** - number of CPUs, CPU frequency in GHz, RAM size in GB, storage size in GB
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS
- ***Example*** - 1 core at 1.7 GHz, 16 GB of RAM, 80 GB of storage

Web Application Capacity Metric

- ***Description*** - measurable characteristics of Web application capacity
- ***Measurement*** - rate of requests per minute
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - SaaS
- ***Example*** - maximum 5.2.0,000 requests per minute

Instance Starting Time Metric

- ***Description*** - length of time required to initialize a new instance
- ***Measurement*** - date/time of instance up - date/time of start request
- ***Frequency*** - per event
- ***Cloud Delivery Model*** - IaaS, PaaS
- ***Example*** - 5 minute maximum, 3 minute average

Response Time Metric

- ***Description*** - time required to perform synchronous operation
- ***Measurement*** - (date/time of request - date/time of response) / total number of requests
- ***Frequency*** - daily, weekly, monthly
- ***Cloud Delivery Model*** - SaaS
- ***Example*** - 5 millisecond average

Completion Time Metric

- ***Description*** - time required to complete an asynchronous task
- ***Measurement*** - (date of request - date of response) / total number of requests
- ***Frequency*** - daily, weekly, monthly
- ***Cloud Delivery Model*** - **PaaS, SaaS**
- **Example** - ***1 second average***

5.2..6.4 Service Scalability Metrics

Service scalability metrics are related to IT resource elasticity capacity, which is related to the maximum capacity that an IT resource can achieve, as well as measurements of its ability to adapt to workload fluctuations. For example, a server can be scaled up to a maximum of 128 CPU cores and 512 GB of RAM, or scaled out to a maximum of 16 load-balanced replicated instances.
The following metrics help determine whether dynamic service demands will be met proactively or reactively, as well as the impacts of manual or automated IT resource allocation processes.

Storage Scalability (Horizontal) Metric

- ***Description*** - permissible storage device capacity changes in response to increased workloads
- ***Measurement*** - storage size in GB
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 1,000 GB maximum (automated scaling)

Server Scalability (Horizontal) Metric

- ***Description*** - permissible server capacity changes in response to increased workloads
- ***Measurement*** - number of virtual servers in resource pool
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS
- ***Example*** - 1 virtual server minimum, 5.2. virtual server maximum (automated scaling)

Server Scalability (Vertical) Metric

- ***Description*** - permissible server capacity fluctuations in response to workload fluctuations
- ***Measurement*** - number of CPUs, RAM size in GB
- ***Frequency*** - continuous
- ***Cloud Delivery Model*** - IaaS, PaaS
- ***Example*** - 512 core maximum, 512 GB of RAM

5.2..6.5 Service Resiliency Metrics

The ability of an IT resource to recover from operational disturbances is often measured using service resiliency metrics. When resiliency is described within or in relation to SLA resiliency guarantees, it is often based on redundant implementations and resource replication over different physical locations, as well as various disaster recovery systems.

The type of cloud delivery model determines how resiliency is implemented and measured. For example, the physical locations of replicated virtual servers that are implementing resilient cloud services can be explicitly expressed in the SLAs for IaaS environments, while being implicitly expressed for the corresponding PaaS and SaaS environments.

Resiliency metrics can be applied in three different phases to address the challenges and events that can threaten the regular level of a service:

- ***Design Phase*** - Metrics that measure how prepared systems and services are to cope with challenges.
- ***Operational Phase*** - Metrics that measure the difference in service levels before, during, and after a downtime event or service outage, which are further qualified by availability, reliability, performance, and scalability metrics.
- ***Recovery Phase*** - Metrics that measure the rate at which an IT resource recovers from downtime, such as the meantime for a system to log an outage and switchover to a new virtual server.

Two common metrics related to measuring resiliency are as follows:

Mean-Time to Switchover (MTSO) Metric

- ***Description*** - the time expected to complete a switchover from a severe failure to a replicated instance in a different geographical area
- ***Measurement*** - (date/time of switchover completion - date/time of failure) / total number of failures
- ***Frequency*** - monthly, yearly
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 5.2. minutes average

Mean-Time System Recovery (MTSR) Metric

- ***Description*** - time expected for a resilient system to perform a complete recovery from a severe failure
- ***Measurement*** - (date/time of recovery - date/time of failure) / total number of failures
- ***Frequency*** - monthly, yearly
- ***Cloud Delivery Model*** - IaaS, PaaS, SaaS
- ***Example*** - 120 minutes average

5.2..7 SLA Guidelines

This section provides a number of best practices and recommendations for working with SLAs, the majority of which are applicable to cloud consumers:

- ***Mapping Business Cases to SLAs*** - It can be helpful to identify the necessary QoS requirements for a given automation solution and to then concretely link them to the guarantees expressed in the SLAs for IT resources responsible for carrying out the automation. This can avoid situations where SLAs are inadvertently misaligned or perhaps unreasonably deviate in their guarantees, subsequent to IT resource usage.
- ***Working with Cloud and On-Premise SLAs*** - Due to the vast infrastructure available to support IT resources in public clouds, the QoS guarantees issued in SLAs for cloud-based IT resources are generally superior to those provided for on-premise IT resources. This variance needs to be understood, especially when building hybrid distributed solutions that utilize both on on-premise and cloud-based services or when incorporating cross-environment technology architectures, such as cloud bursting.
- ***Understanding the Scope of an SLA*** - Cloud environments are comprised of many supporting architectural and infrastructure layers upon which IT resources reside and are integrated. It is important to acknowledge the extent to which a given IT resource guarantee applies. For example, an SLA may be limited to the IT resource implementation but not its underlying hosting environment.
- ***Understanding the Scope of SLA Monitoring*** - SLAs need to specify where monitoring is performed and where measurements are calculated, primarily in relation to the cloud's firewall. For example, monitoring within the cloud firewall is not always advantageous or relevant to the cloud consumer's required QoS guarantees. Even the most efficient firewalls have a measurable degree of influence on performance and can further present a

point of failure.

- ***Documenting Guarantees at Appropriate Granularity*** - SLA templates used by cloud providers sometimes define guarantees in broad terms. If a cloud consumer has specific requirements, the corresponding level of detail should be used to describe the guarantees. For example, if data replication needs to take place across particular geographic locations, then these need to be specified directly within the SLA.
- ***Defining Penalties for Non-Compliance*** - If a cloud provider is unable to follow through on the QoS guarantees promised within the SLAs, recourse can be formally documented in terms of compensation, penalties, reimbursements, or otherwise.
- ***Incorporating Non-Measurable Requirements*** - Some guarantees cannot be easily measured using service quality metrics, but are relevant to QoS nonetheless, and should therefore still be documented within the SLA. For example, a cloud consumer may have specific security and privacy requirements for data hosted by the cloud provider that can be addressed by assurances in the SLA for the cloud storage device being leased.
- ***Disclosure of Compliance Verification and Management*** - Cloud providers are often responsible for monitoring IT resources to ensure compliance with their own SLAs. In this case, the SLAs themselves should state what tools and practices are being used to carry out the compliance checking process, in addition to any legal-related auditing that may be occurring.
- ***Inclusion of Specific Metric Formulas*** - Some cloud providers do not mention common SLA metrics or the metrics-related calculations in their SLAs, instead focusing on service-level descriptions that highlight the use of best practices and customer support. Metrics being used to measure SLAs should be part of the SLA document, including the formulas and calculations that the metrics are based upon.
- ***Considering Independent SLA Monitoring*** - Although cloud providers will often have sophisticated SLA management systems and SLA monitors, it may be in the best interest of a cloud consumer to hire a third-party organization to perform independent monitoring as well, especially if there are suspicions that SLA guarantees are not always being met by the cloud provider (despite the results shown on periodically issued monitoring reports).
- ***Archiving SLA Data*** - The SLA-related statistics collected by SLA monitors are commonly stored and archived by the cloud provider for future reporting purposes. If a cloud provider intends to keep SLA data specific to a cloud consumer even after the cloud consumer no longer continues its business relationship with the cloud provider, then this should be disclosed. The cloud consumer may have data privacy requirements that disallow the unauthorized storage of this type of information. Similarly, during and after a cloud consumer's engagement with a cloud provider, it may want to keep a copy of historical SLA-related data as well. It may be especially useful for comparing cloud providers in the future.
- ***Disclosing Cross-Cloud Dependencies*** - Cloud providers may be leasing IT resources from other cloud providers, which results in a loss of control over the guarantees they are able to make to cloud consumers. Although a cloud provider will rely on the SLA assurances made to it by other cloud providers, the cloud consumer may want disclosure of the fact that

the IT resources it is leasing may have dependencies beyond the environment of the cloud provider organization.