Cloud computing

Information Storage And Management (SRM Institute of Science and Technology)

Chapter 13
Cloud Computing

KEY CONCEPTS

- Essential Characteristics of Cloud Computing
- Cloud Services and Deployment Models
- Cloud Computing Infrastructure
- Cloud Adoption Considerations

In today's competitive environment, organizations are under increasing pressure to improve efficiency and transform their IT processes to achieve more with less. Businesses need reduced time-to-market, better agility, higher availability, and reduced expenditures to meet the changing business requirements and accelerated pace of innovation. These business requirements are posing several challenges to IT teams. Some of the key challenges are serving customers worldwide around the clock, refreshing technology quickly and faster provisioning of IT resources — all at reduced costs.

These long-standing challenges are addressed with the emergence of a new computing style, called *cloud computing*, which enables organizations and individuals to obtain and provision IT resources as a service. With cloud computing, users can browse and select relevant cloud services, such as compute, software, storage, or a combination of these resources, via a portal. Cloud computing automates delivery of selected cloud services to the users. It helps organizations and individuals deploy IT resources at reduced total cost of ownership with faster provisioning and compliance adherence.

A widely adopted definition of cloud computing comes from the U.S. National Institute of Standards and Technology (NIST Special Publication 800-145):

> *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.*

This chapter covers the enabling technologies, essential characteristics, benefits, services, deployment models, and infrastructure of cloud computing. The chapter also includes the challenges and considerations in adopting cloud computing.

13.1 Cloud Enabling Technologies

Grid computing, utility computing, virtualization, and service-oriented architecture are enabling technologies of cloud computing.

- *Grid computing* is a form of distributed computing that enables the resources of numerous heterogeneous computers in a network to work together on a single task at the same time. Grid computing enables parallel computing and is best for large workloads.
- *Utility computing* is a service-provisioning model in which a service provider makes computing resources available to customers, as required, and charges them based on usage. This is analogous to other utility services, such as electricity, where charges are based on the consumption.
- *Virtualization* is a technique that abstracts the physical characteristics of IT resources from resource users. It enables the resources to be viewed and managed as a pool and lets users create virtual resources from the pool. Virtualization provides better flexibility for provisioning of IT resources compared to provisioning in a non-virtualized environment. It helps optimize resource utilization and delivering resources more efficiently.
- *Service Oriented Architecture* (SOA) provides a set of services that can communicate with each other. These services work together to perform some activity or simply pass data among services.

13.2 Characteristics of Cloud Computing

A computing infrastructure used for cloud services must have certain capabilities or characteristics. According to NIST, the cloud infrastructure should have five essential characteristics:

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

  A cloud service provider publishes a service catalogue, which contains information about all cloud services available to consumers. The service catalogue includes information about service attributes, prices, and request processes. Consumers view the service catalogue via a web-based user

interface and use it to request for a service. Consumers can either leverage the "ready-to-use" services or change a few service parameters to customize the services.

- **Broad network access:** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (for example, mobile phones, tablets, laptops, and workstations).
- **Resource pooling:** The provider's computing resources are pooled to serve multiple consumers using a multitenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (for example, country, state, or data center). Examples of resources include storage, processing, memory, and network bandwidth.
- **Rapid elasticity:** Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

  Consumers can leverage rapid elasticity of the cloud when they have a fluctuation in their IT resource requirements. For example, an organization might require double the number of web and application servers for a specific duration to accomplish a specific task. For the remaining period, they might want to release idle server resources to cut down the expenses. The cloud enables consumers to grow and shrink the demand for resources dynamically.
- **Measured service:** Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (for example, storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

MULTITENANCY

Multitenancy **refers to an architecture in which multiple independent consumers (tenants) are serviced using a single set of resources. This lowers the cost of services for consumers. Virtualization enables resource pooling and multitenancy in the cloud. For example, multiple virtual machines from different consumers can run simultaneously on the same physical server that runs the hypervisor.**

13.3 Benefits of Cloud Computing

Cloud computing offers the following key benefits:

- **Reduced IT cost:** Cloud services can be purchased based on pay-per-use or subscription pricing. This reduces or eliminates the consumer's IT capital expenditure (CAPEX).
- **Business agility:** Cloud computing provides the capability to allocate and scale computing capacity quickly. Cloud computing can reduce the time required to provision and deploy new applications and services from months to minutes. This enables businesses to respond more quickly to market changes and reduce time-to-market.
- **Flexible scaling:** Cloud computing enables consumers to scale up, scale down, scale out, or scale in the demand for computing resources easily. Consumers can unilaterally and automatically scale computing resources without any interaction with cloud service providers. The flexible service provisioning capability of cloud computing often provides a sense of unlimited scalability to the cloud service consumers.
- **High availability:** Cloud computing has the capability to ensure resource availability at varying levels depending on the consumer's policy and priority. Redundant infrastructure components (servers, network paths, and storage equipment, along with clustered software) enable fault tolerance for cloud deployments. These techniques can encompass multiple data centers located in different geographic regions, which prevents data unavailability due to regional failures.

13.4 Cloud Service Models

According to NIST, cloud service offerings are classified primarily into three models: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

13.4.1 Infrastructure-as-a-Service

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems and deployed applications; and possibly limited control of select networking components (for example, host firewalls).

IaaS is the base layer of the cloud services stack (see Figure 13-1 [a]). It serves as the foundation for both the SaaS and PaaS layers.

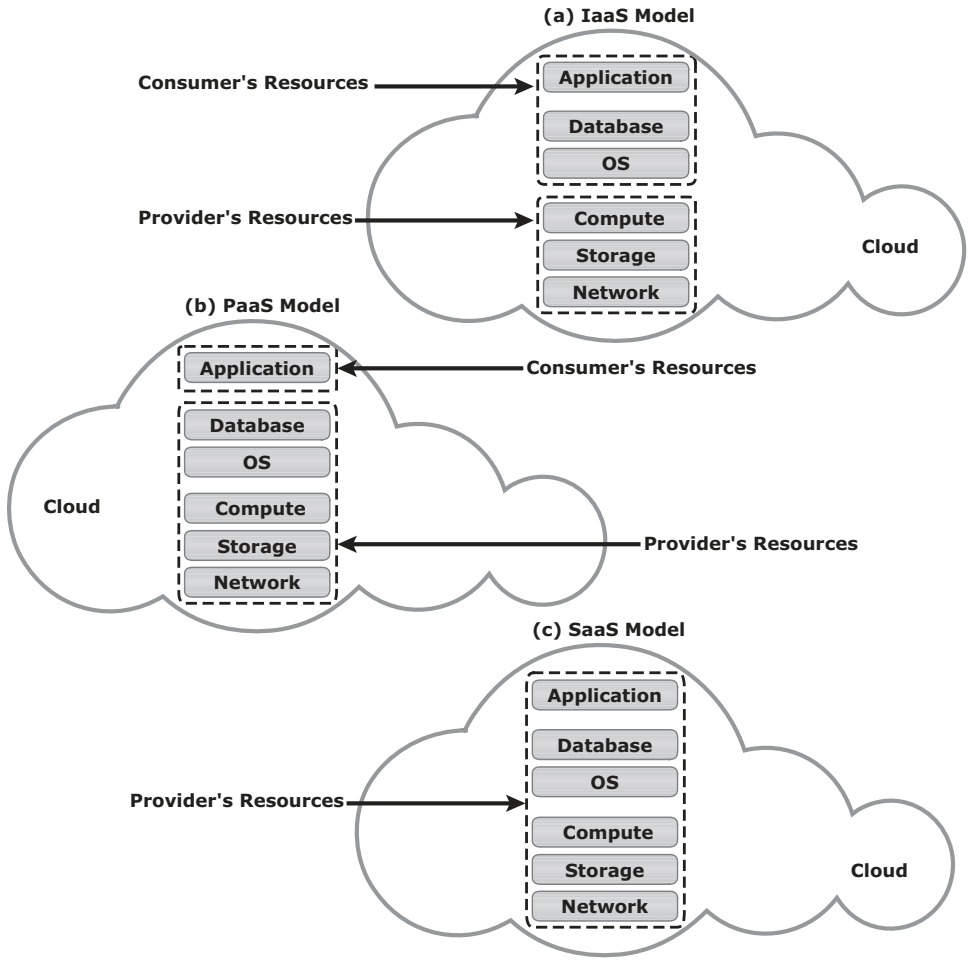


Figure 13-1: IaaS, PaaS, and SaaS models

Amazon Elastic Compute Cloud (Amazon EC2) is an example of IaaS that provides scalable compute capacity, on-demand, in the cloud. It enables consumers to leverage Amazon's massive computing infrastructure with no up-front capital investment.

13.4.2 Platform-as-a-Service

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not

manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. (See Figure 13-1 [b]).

PaaS is also used as an application development environment, offered as a service by the cloud service provider. The consumer may use these platforms to code their applications and then deploy the applications on the cloud. Because the workload to the deployed applications varies, the scalability of computing resources is usually guaranteed by the computing platform, transparently. Google App Engine and Microsoft Windows Azure Platform are examples of PaaS.

13.4.3 Software-as-a-Service

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (for example, web-based e-mail), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. (See Figure 13-1[c]).

In a SaaS model, applications, such as customer relationship management (CRM), e-mail, and instant messaging (IM), are offered as a service by the cloud service providers. The cloud service providers exclusively manage the required computing infrastructure and software to support these services. The consumers may be allowed to change a few application configuration settings to customize the applications.

EMC Mozy is an example of SaaS. Consumers can leverage the Mozy console to perform automatic, secured, online backup and recovery of their data with ease. Salesforce.com is a provider of SaaS-based CRM applications, such as Sales Cloud and Service Cloud.

13.5 Cloud Deployment Models

According to NIST, cloud computing is classified into four deployment models — public, private, community, and hybrid — which provide the basis for how cloud infrastructures are constructed and consumed.

13.5.1 Public Cloud

In a *public cloud* model, the cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Consumers use the cloud services offered by the providers via the Internet and pay metered usage charges or subscription fees. An advantage of the public cloud is its low capital cost with enormous scalability. However, for consumers, these benefits come with certain risks: no control over the resources in the cloud, the security of confidential data, network performance, and interoperability issues. Popular public cloud service providers are Amazon, Google, and Salesforce.com. Figure 13-2 shows a public cloud that provides cloud services to organizations and individuals.

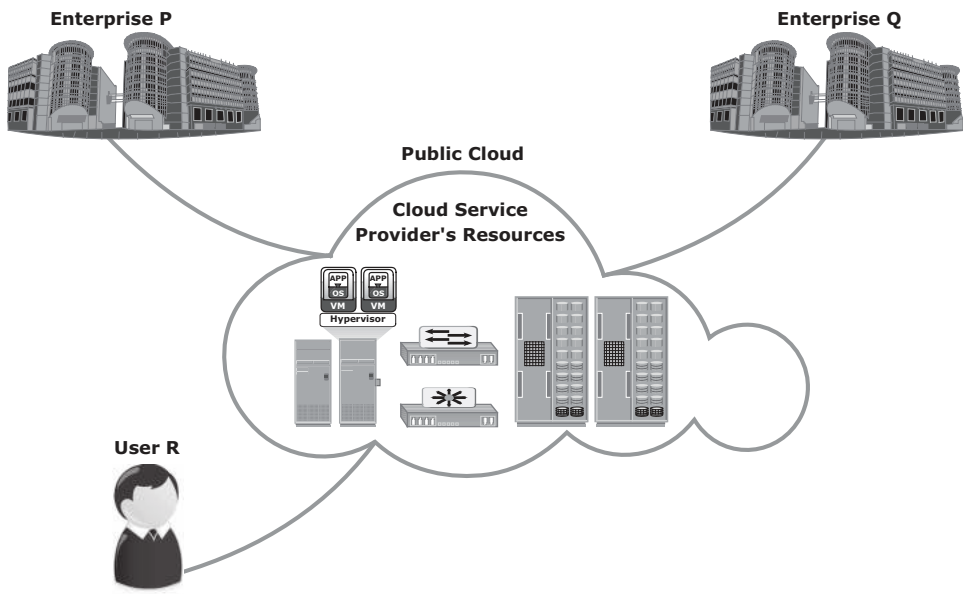
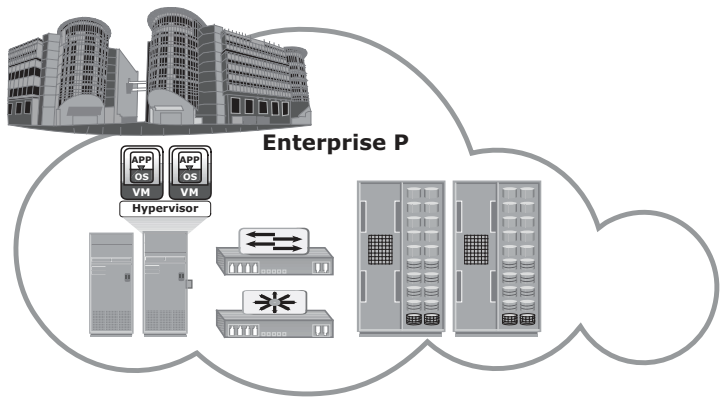


Figure 13-2: Public cloud

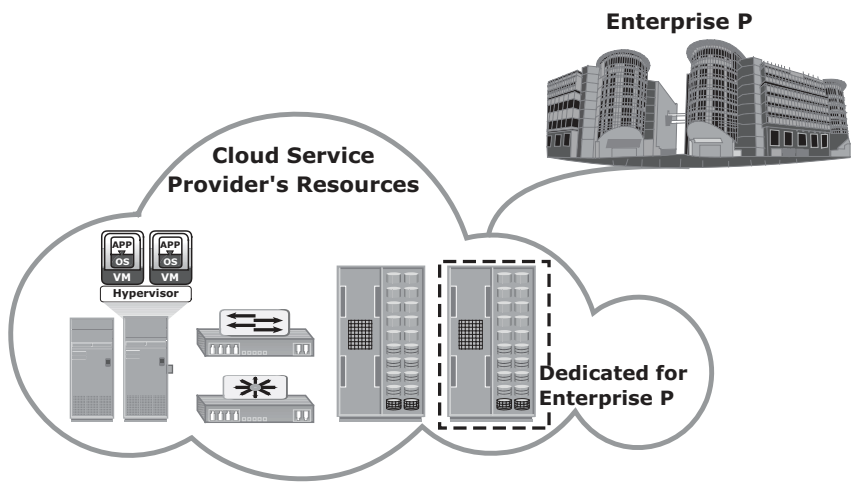
13.5.2 Private Cloud

In a *private cloud* model, the cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (for example, business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises. Following are two variations to the private cloud model:

- **On-premise private cloud:** The on-premise private cloud, also known as internal cloud, is hosted by an organization within its own data centers (see Figure 13-3 [a]). This model enables organizations to standardize their cloud service management processes and security, although this model has limitations in terms of size and resource scalability. Organizations would also need to incur the capital and operational costs for the physical resources. This is best suited for organizations that require complete control over their applications, infrastructure configurations, and security mechanisms.



(a) On-Premise Private Cloud



(b) Externally Hosted Private Cloud

Figure 13-3: On-premise and externally hosted private clouds

- **Externally hosted private cloud:** This type of private cloud is hosted external to an organization (see Figure 13-3 [b]) and is managed by a third-party organization. The third-party organization facilitates an exclusive cloud environment for a specific organization with full guarantee of privacy and confidentiality.

13.5.3 Community Cloud

In a *community cloud* model, the cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared

concerns (for example, mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises. (See Figure 13-4).

In a community cloud, the costs spread over to fewer consumers than a public cloud. Hence, this option is more expensive but might offer a higher level of privacy, security, and compliance. The community cloud also offers organizations access to a vast pool of resources compared to the private cloud. An example in which a community cloud could be useful is government agencies. If various agencies within the government operate under similar guidelines, they could all share the same infrastructure and lower their individual agency's investment.

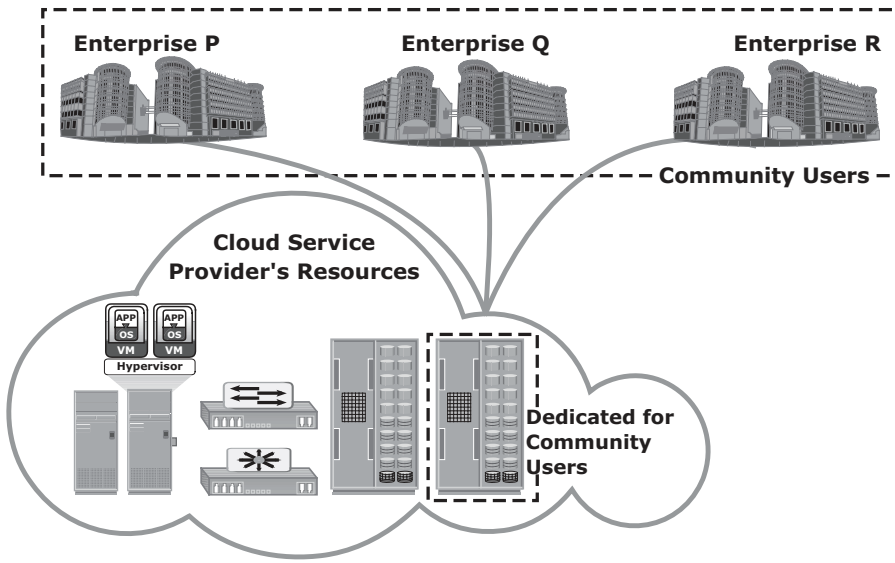


Figure 13-4: Community cloud

13.5.4 Hybrid Cloud

In a *hybrid cloud* model, the cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (for example, cloud bursting for load balancing between clouds).

The hybrid model allows an organization to deploy less critical applications and data to the public cloud, leveraging the scalability and cost-effectiveness of the public cloud. The organization's mission-critical applications and data remain on the private cloud that provides greater security. Figure 13-5 shows an example of a hybrid cloud.

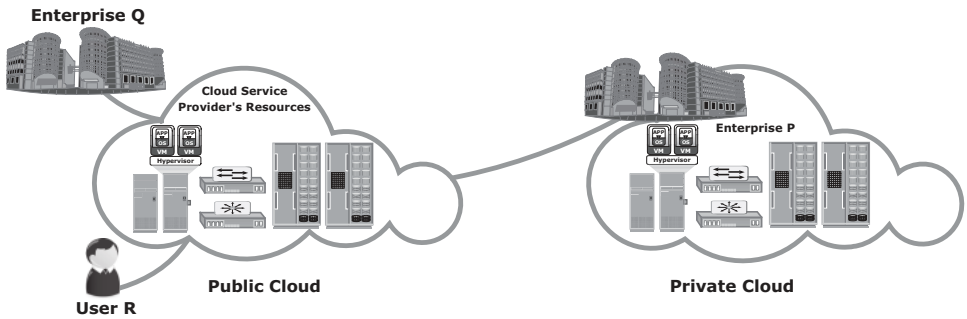


Figure 13-5: Hybrid cloud

13.6 Cloud Computing Infrastructure

A cloud computing infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing. Cloud computing infrastructure usually consists of the following layers:

- Physical infrastructure
- Virtual infrastructure
- Applications and platform software
- Cloud management and service creation tools

The resources of these layers are aggregated and coordinated to provide cloud services to the consumers (see Figure 13-6).

13.6.1 Physical Infrastructure

The physical infrastructure consists of physical computing resources, which include physical servers, storage systems, and networks. Physical servers are connected to each other, to the storage systems, and to the clients via networks, such as IP, FC SAN, IP SAN, or FCoE networks.

Cloud service providers may use physical computing resources from one or more data centers to provide services. If the computing resources are distributed across multiple data centers, connectivity must be established among them. The connectivity enables the data centers in different locations to work as a single large data center. This enables migration of business applications and data across data centers and provisioning cloud services using the resources from multiple data centers.