High-Level View of Cloud Computing

NIST Definition and Characteristics of Cloud Computing

NIST defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction".

According to NIST, there are **five essential** characteristics of cloud computing:

1. **On-demand self-service:** Consumers can acquire resources based on service demand without human intervention.
2. **Broad network access:** Cloud services are accessible remotely from diverse client platforms.
3. **Resource pooling:** Resources are shared among consumers in a multi-tenant fashion.
4. **Rapid elasticity:** Cloud resources can be provisioned and released quickly, often in real-time, with minimal human involvement, giving the illusion of infinite resources.
5. **Measured service:** Resource usage is controlled, monitored, and reported (often for billing purposes, e.g., pay-per-use).

The key characteristic is **elasticity**, which manifests in two dimensions:

- **On-demand in time:** Users can change resource requests on short notice, potentially for short periods.
- **On-demand in scale:** Users can start with minimal resources and expand to very large scales. For example, Target, a major retail chain, runs its web and inventory control systems in a rented cloud to leverage this elasticity.

Virtualization

Virtualization is a core enabling technology for cloud computing. In a traditional stack, applications run on an operating system, which runs directly on hardware. In a virtualized stack, a hypervisor sits between the hardware and multiple operating systems, allowing multiple applications to run concurrently on different operating systems, all sharing the same underlying hardware. This technology creates a pool of storage and computing resources by partitioning physical resources.

Resources Provided as a Service

Cloud computing leverages a service-driven business model, providing hardware and platform-level resources as services on an on-demand basis. Users can request, configure, and access cloud resources using cloud-specific APIs. This enables consumers to obtain computing resources as and when needed, without human intervention, and choose services from a catalog.

Cloud computing services are typically categorized into three main models:

1. **Infrastructure-as-a-Service (IaaS):**

    - Provides the capability to the consumer to hire infrastructure components such as servers, storage, and network.
    - Consumers can deploy and run software, including operating systems and applications, having control over virtual resources.

    - Examples include Amazon Web Services (AWS), VirtualBox, VMware, and OpenStack. IaaS is responsible for managing the physical resources, such as servers, routers, switches, and power/cooling systems, typically implemented in data centers.

2. **Platform-as-a-Service (PaaS):**

    - Offers the capability to deploy consumer-created or acquired applications on the cloud provider's infrastructure.
    - The consumer has control over deployed applications and their hosting environment configurations.
    - Billing is typically based on platform software components, and clients may use proprietary languages.
    - Examples include Google App Engine, Microsoft Azure, and Facebook platform. PaaS consists of operating systems and application frameworks, aiming to minimize the burden of deploying applications directly into VM containers.

3. **Software-as-a-Service (SaaS):**

    - Provides the capability to use the provider's applications running in a cloud infrastructure.
    - The complete stack, including the application, is provided as a service.
    - Applications are accessible from various client devices, such as web browsers (thin client interface).
    - Billing is based on application usage.
    - Examples include Dropbox, Google Docs, Google Apps of G.suite, web email, and Salesforce.com.

Cloud Computing Service Categories

A visual representation shows the different layers of management responsibility for IaaS, PaaS, and SaaS.

- **IaaS:** The user manages the application, data, runtime, middleware, and operating system. The vendor manages virtualization, servers, storage, and networking.
- **PaaS:** The user manages the application and data. The vendor manages runtime, middleware, operating system, virtualization, servers, storage, and networking.
- **SaaS:** The vendor manages the entire stack: application, data, runtime, middleware, operating system, virtualization, servers, storage, and networking.

Pay-Per-Use Business Model

In the pay-per-use model, consumers only pay for the resources they actually use. Resource usage is monitored and reported, providing transparency for charge-back to both the cloud service provider and the consumer. Pricing and billing models are linked to the required service levels.

An example is Amazon EC2 (Elastic Compute Cloud) for general purpose instances, where pricing is per hour based on virtual CPU (VCPU), Memory (GB), and Instance Storage (GB). For example, a `t2.nano` instance costs $0.0075 per hour with 1 VCPU and 0.5 GB memory, while an `m4.10xlarge` costs $2.85 per hour with 40 VCPUs and 160 GB memory. EBS stands for Elastic Block Store.

Elasticity

Elasticity is a core characteristic where consumers can acquire or release resources on demand. It refers to the ability to rapidly scale IT resources to fulfill changing needs without service interruption, allowing resources to be scaled both up and down dynamically. To the consumer, the cloud appears to offer infinite capacity. Consumers can start with minimal computing power and expand their environment to any size.

Elasticity helps in avoiding over-provisioning (which leads to underutilization) and under-provisioning (which leads to lost revenue or users). Data centers are expensive, costing over $150 million and taking 24+ months to design and build. The pay-by-use model allows users to align capacity with demand, rather than provisioning for peak demand, which would otherwise lead to wasted resources. Conversely, under-provisioning means capacity falls below demand, resulting in lost revenue or users.

Enabling Technologies

Several technologies enable cloud computing:

- **Data Centers:** The physical infrastructure housing computing resources.
- **Machine Virtualization:** Allows multiple virtual machines to run on a single physical server.
- **Networking:** Crucial for inter-server communication within the data center and external connectivity.
- **Data Storage and Management:** Involves distributed storage systems that manage data consistency and availability despite failures.
- **Distributed Processing:** Paradigms like Map/Reduce facilitate large-scale data processing.
- **Resource Management and Scheduling:** Mechanisms to efficiently allocate and manage computing and networking resources.
- **Energy Management:** Focuses on optimizing power consumption and cooling within data centers.
- **Security and Privacy:** Addresses concerns related to data protection and access control in the cloud environment.