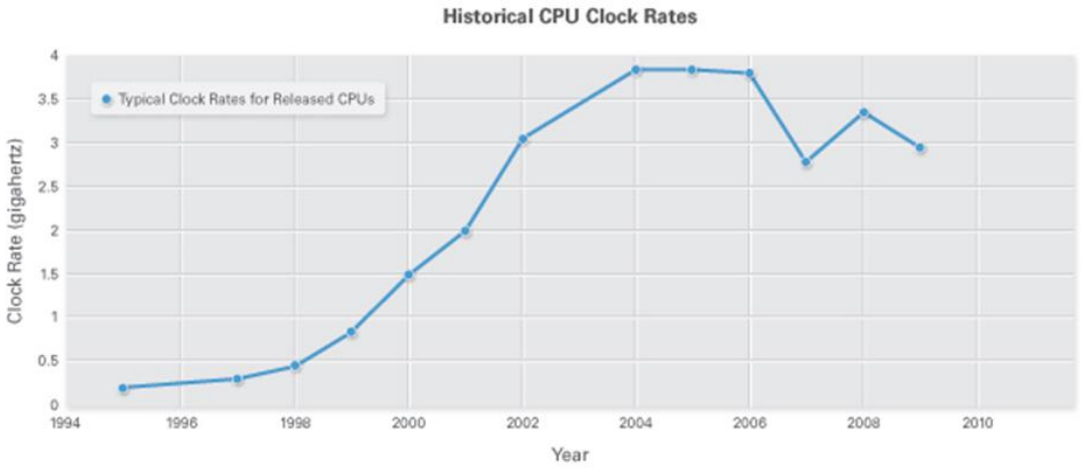
MODULE 1

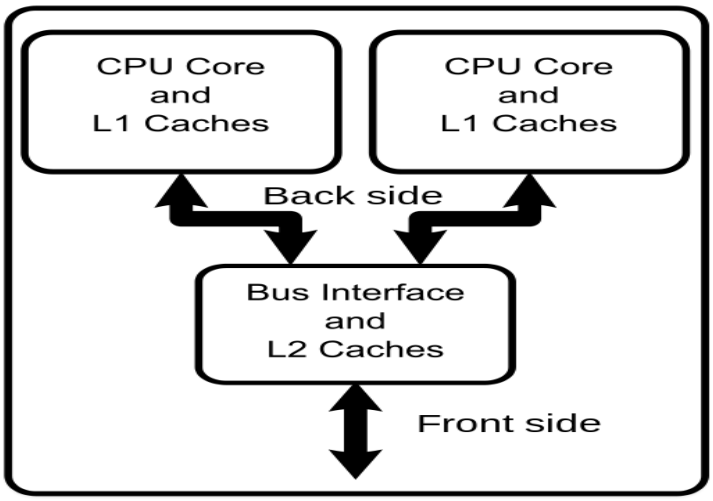
CLOUD COMPUTING

TECHNOLOGIES FOR NETWORK-BASED SYSTEMS

1. Multi-core CPUs and Multithreading Technologies

CPUs today assume a multi-core architecture with dual, quad, six, or more processing cores. The clock rate increased from 10 MHz for Intel 286 to 4 GHz for Pentium 4 in 30 years. However, the clock rate reached its limit on CMOS chips due to power limitations. Clock speeds cannot continue to increase due to excessive heat generation and current leakage.





LI cache is private to each core, L2 cache is shared and L3 cache or DRAM is off the chip. Examples of multi-core CPUs include Intel i7, Xeon, AMD Opteron. Each core can also be multithreaded. E.g., the Niagara II has 8 cores with each core handling 8 threads for a total of 64 threads maximum.

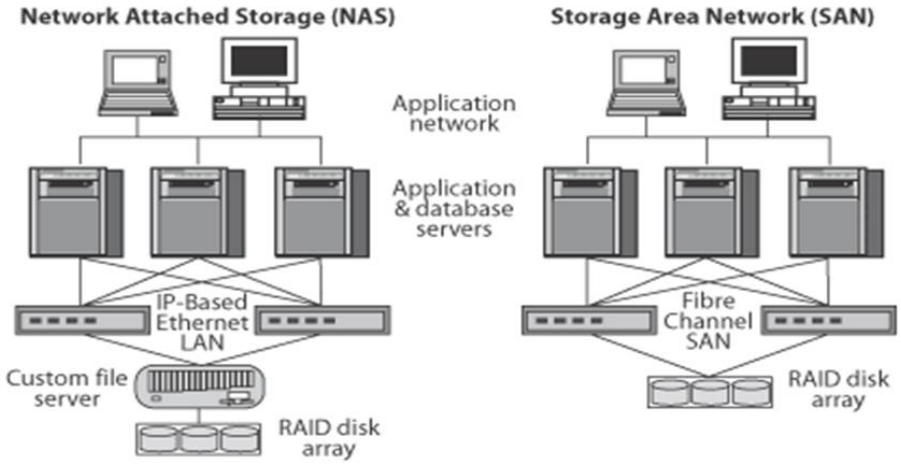
2. Memory, Storage and WAN

- DRAM chip capacity increased from 16 KB in 1976 to 64 GB in 2011 for a 4x increase in capacity every 3 years. Memory access time did not improve as much.
- For hard drives, capacity increased from 260 MB in 1981 to 3 TB for the Seagate Barracuda XT hard drive in 2011 for an approximate 10x increase in capacity every 8 years.
- The "memory wall" is the growing disparity of speed between CPU and memory outside the CPU chip. An important reason for this disparity is the limited communication bandwidth beyond chip boundaries. From 1986 to 2000, CPU speed improved at an annual rate of 55% while memory speed only improved at 10%.
- Faster processor speed and larger memory capacity result in a wider performance gap between processors and memory. The memory wall may become an even worse problem limiting CPU performance.

System-Area Interconnects

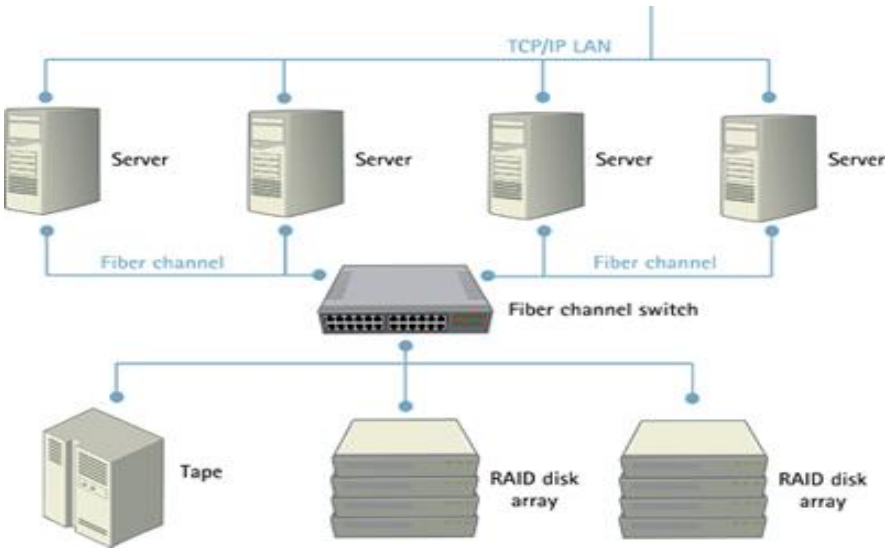
A LAN is typically used to connect clients to big servers. A Storage Area Network (SAN) connects servers to network storage such as disk arrays. Network attached storage (NAS) connects servers directly to disk arrays. All 3 types of networks often appear in a large cluster built with commercial network components.

A NAS is fundamentally a bunch of disks, usually arranged in a disk array. Users and servers attach to the NAS primarily using TCP/IP over Ethernet, and the NAS has its own IP address. The primary job of a NAS is to serve files, so most NAS systems offer support for Windows networking, HTTP, plus file systems and protocols such as NFS.



One way to loosely conceptualize the difference between a NAS and a SAN is that a NAS appears to the client OS (operating system) as a file server (the client can map network drives to shares on that server) whereas a disk available through a SAN still appears to the client OS as a disk, visible in disk and volume management utilities (along with client's local disks), and available to be formatted with a file system and mounted.

SANs allow multiple servers to share a RAID, making it appear to the server as if it were local or directly attached storage, and it cannot be accessed by individual users. A dedicated networking standard, Fiber Channel, allow blocks to be moved between servers and storage at high speed. It uses dedicated switches and a fiber-based cabling system which separates it from the day-to-day traffic. It uses the SCSI protocol for communication.



3. Virtual Machines and Virtualization Middleware

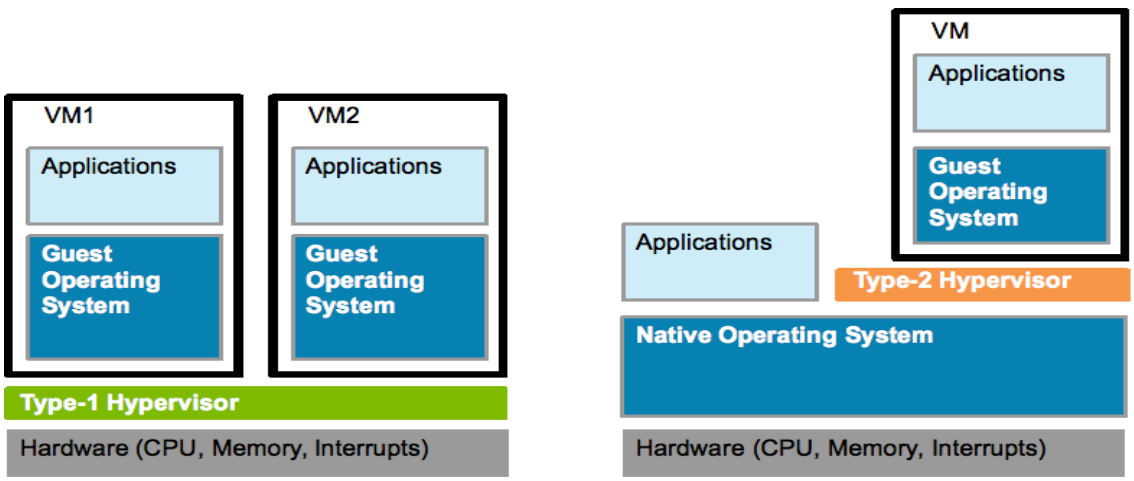
To build clouds we need to aggregate large amounts of computing, storage, and networking resources in a virtualized manner. Specifically, clouds rely on the dynamic virtualization of CPU, memory, and I/O.

Virtual Machines (VMs)

- The VM is built with virtual resources managed by a guest OS to run a specific application.
- Between the VMs and the host platform, a middleware layer (called the Virtual Memory Monitor (VMM) or a hypervisor) is deployed.

Type 1 (bare metal) hypervisor

- The bare metal hypervisor runs on the bare hardware and handles all the Hardware (CPU, memory, and I/O) directly. This runs in the privileged mode.
- The guest OS could any OS such as Linux, Windows etc.
- They provide an almost native performance to the guest OSs (VMs), generally losing only 3–4% of the Central Processing Unit's cycles to the running of the hypervisor.
- Bare-metal is great for consolidating a company's collection of servers onto a single piece of hardware.
- Some examples of the leading bare-metal hypervisors are VMware's ESX(i) (proprietary), Citrix Xen Server (FOS (Free & Open Source)), and KVM (kernel loaded VM) (FOS). ESX and XenServer are installs that reside directly on the hardware. KVM sits within a Linux kernel.



Type 2 (or hosted) hypervisor

- Here the hypervisor runs on top of a host OS in user or non-privileged mode.
- The host OS need not be modified. For example, you could install Windows or Linux, and then install the hypervisor on top but the performance may not be as good as with bare-metal.
- Hosted is often used by IT workers who need the flexibility to install, run and try out different OSs on their own computers without disrupting their current computing environment.
- Some examples of the leading hosted hypervisors are VirtualBox (FOS) and Qemu (FOS)
- Many VMs can be run on a hypervisor. The resource most in demand is system memory, and because RAM is cheap, this makes the proposition of virtualization an attractive one.
- A VM can be suspended and stored in secondary storage, resumed, or migrated from one hardware platform to another.

Full Virtualization vs. Para-Virtualization

- Full Virtualization allows the guest OS to run on the hypervisor without any modification and without it knowing that it is hosted.
- Paravirtualization requires that the kernel of the guest OS is modified and compiled with hooks (an API) for the hypervisor (guest OSs must know about the hypervisor).
- The guest OS can then communicate and cooperates with the hypervisor with a potential to improving performance, though this might be marginal (load that generates system calls benefits the most).
- Windows guests can only run-on Full Virtualization, as their source is proprietary.
- Operating systems that support paravirtualization interfaces need custom kernel adjustments.
- If compiled manually (where possible), guest operating systems with hypervisor support require more maintenance and configuration.
- These additional costs and complexity, combined with the marginal performance gains, means paravirtualization remains a niche product in the server virtualization market.

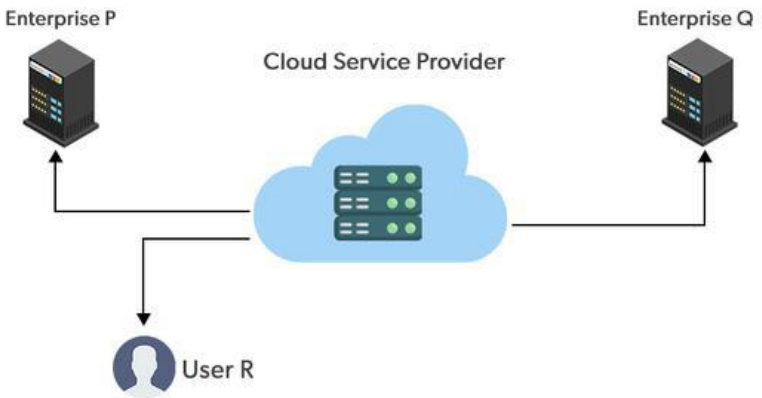
TYPES OF CLOUD

Cloud computing is Internet-based computing in which a shared pool of resources is available over broad network access. These resources can be provisioned or released with minimum management efforts and service provider interaction.

1. Public cloud
2. Private cloud
3. Hybrid cloud
4. Community cloud

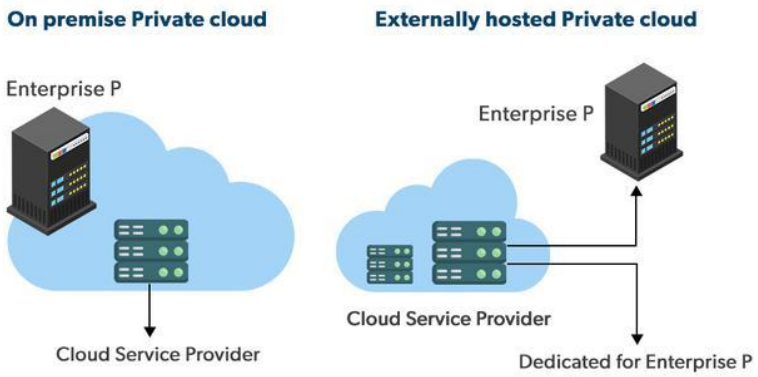
Public Cloud

- Public clouds are managed by third parties which provide cloud services over the internet to the public.
- These services are available as pay-as-you-go billing models.
- They offer solutions for minimizing IT infrastructure costs and become a good option for handling peak loads on the local infrastructure.
- Public clouds are the go-to option for small enterprises, which are able to start their businesses without large upfront investments by completely relying on public infrastructure for their IT needs.
- The fundamental characteristics of public clouds are multitenancy.
- A public cloud is meant to serve multiple users, not a single customer.
- A user requires a virtual computing environment that is separated, and most likely isolated, from other users.



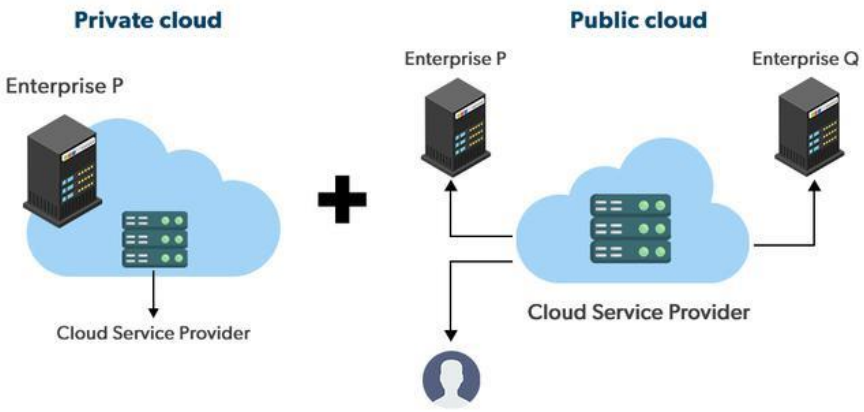
Private cloud

- Private clouds are distributed systems that work on private infrastructure.
- Provide the users with dynamic provisioning of computing resources.
- Instead of a pay-as-you-go model in private clouds, there could be other schemes that manage the usage of the cloud and proportionally billing of the different departments or sections of an enterprise.



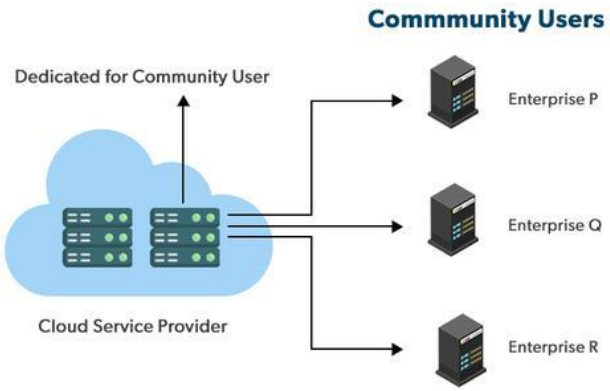
Hybrid cloud

- A hybrid cloud is a heterogeneous distributed system formed by combining facilities of public cloud and private cloud.
- For this reason, they are also called heterogeneous clouds.
- A major drawback of private deployments is the inability to scale on-demand and efficiently address peak loads.
- Here public clouds are needed.
- Hence, a hybrid cloud takes advantage of both public and private clouds.



Community cloud

- Community clouds are distributed systems created by integrating the services of different clouds to address the specific needs of an industry, a community, or a business sector.
- In the community cloud, the infrastructure is shared between organizations that have shared concerns or tasks.
- The cloud may be managed by an organization or a third party.



- **Sectors that use community clouds**
- **Media industry:**
  - ✓ Media companies are looking for quick, simple, low-cost ways for increasing the efficiency of content generation.
  - ✓ Most media productions involve an extended ecosystem of partners.
  - ✓ In particular, the creation of digital content is the outcome of a collaborative process that includes the movement of large data, massive compute-intensive rendering tasks, and complex workflow executions.
- **Healthcare industry:**
  - ✓ In the healthcare industry community clouds are used to share information and knowledge on the global level with sensitive data in the private infrastructure.
- **Energy and core industry:**
  - ✓ In these sectors, the community cloud is used to cluster a set of solution which collectively addresses management, deployment, and orchestration of services and operations.

- **Scientific research:**
  - ✓ In this, organization with common interests in science share a large distributed infrastructure for scientific computing.

CLOUD SERVICE MODELS

There are the following three types of cloud service models

1. Infrastructure as a Service (IaaS)
2. Platform as a Service (PaaS)
3. Software as a Service (SaaS)

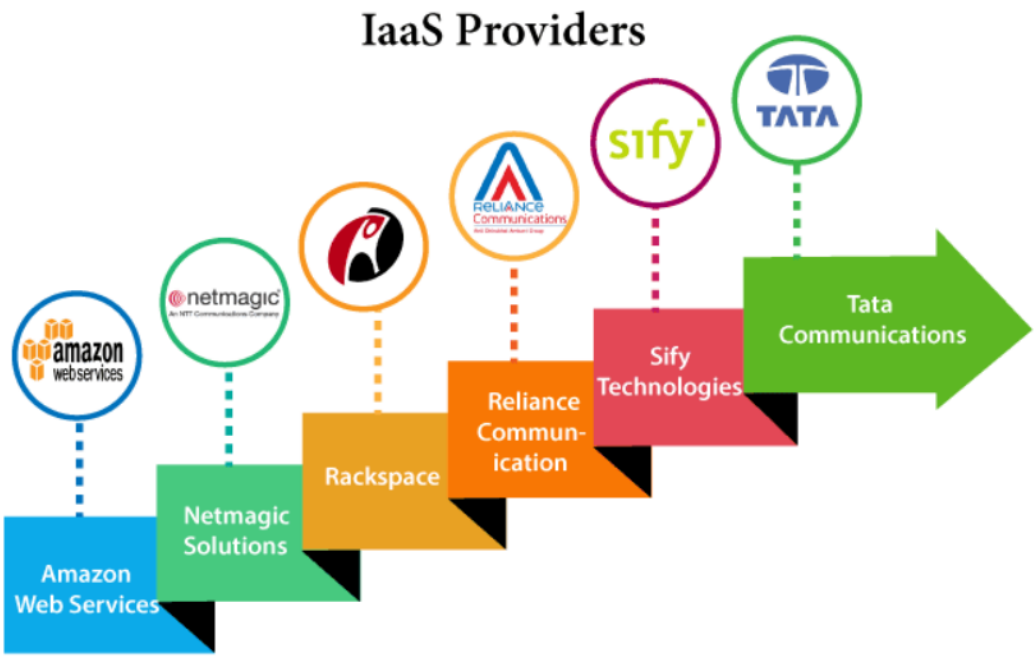
Infrastructure as a Service (IaaS)

IaaS is also known as Hardware as a Service (HaaS). It is one of the layers of the cloud computing platform. It allows customers to outsource their IT infrastructures such as servers, networking, processing, storage, virtual machines, and other resources. Customers access these resources on the Internet using a pay-as-per use model. In traditional hosting services, IT infrastructure was rented out for a specific period of time, with pre-determined hardware configuration.

The client paid for the configuration and time, regardless of the actual use. With the help of the IaaS cloud computing platform layer, clients can dynamically scale the configuration to meet changing requirements and are billed only for the services actually used. IaaS is offered in three models: public, private, and hybrid cloud. The private cloud implies that the infrastructure resides at the customer-premise. In the case of public cloud, it is located at the cloud computing platform vendor's data center. Hybrid cloud is a combination of the two in which the customer selects the best of both public cloud or private cloud.

IaaS provider provides the following services:

1. Compute: Computing as a Service includes virtual central processing units and virtual main memory for the Vms that is provisioned to the end- users.
2. Storage: IaaS provider provides back-end storage for storing files.
3. Network: Network as a Service (NaaS) provides networking components such as routers, switches, and bridges for the Vms.
4. Load balancers: It provides load balancing capability at the infrastructure layer.



Platform as a Service (PaaS)

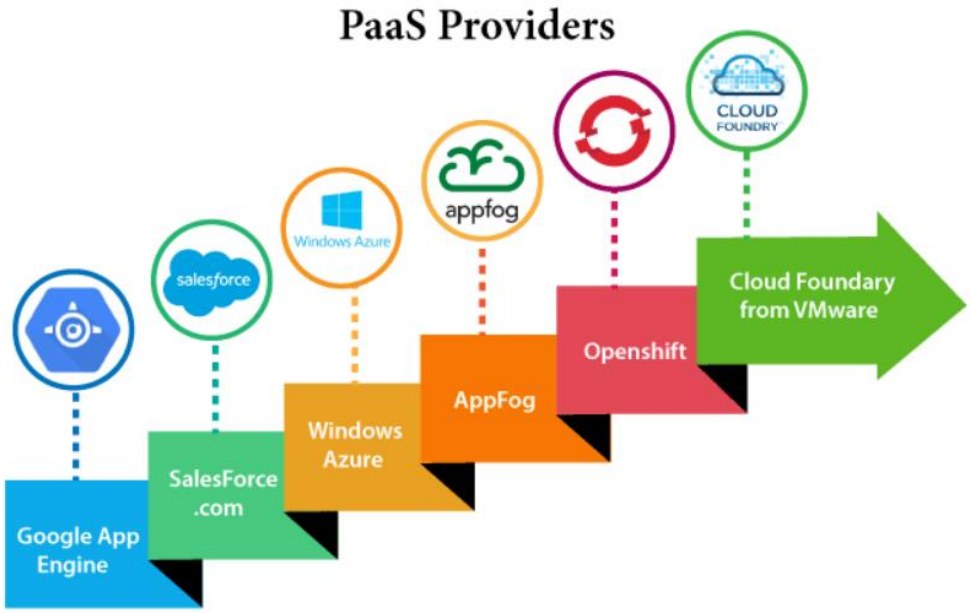
Platform as a Service (PaaS) provides a runtime environment. It allows programmers to easily create, test, run, and deploy web applications. You can purchase these applications from a cloud service provider on a pay-as-per use basis and access them using the Internet connection. In PaaS, back-end scalability is managed by the cloud service provider, so end- users do not need to worry about managing the infrastructure.

PaaS includes infrastructure (servers, storage, and networking) and platform (middleware, development tools, database management systems, business intelligence, and more) to support the web application life cycle.

PaaS providers provides the following services:

1. Programming languages: PaaS providers provide various programming languages for the developers to develop the applications. Some popular programming languages provided by PaaS providers are Java, PHP, Ruby, Perl, and Go.
2. Application frameworks: PaaS providers provide application frameworks to easily understand the application development. Some popular application frameworks provided by PaaS providers are Node.js, Drupal, Joomla, WordPress, Spring, Play, Rack, and Zend.

3. Databases: PaaS providers provide various databases such as ClearDB, PostgreSQL, MongoDB, and Redis to communicate with the applications.
4. Other tools: PaaS providers provide various other tools that are required to develop, test, and deploy the applications.



Software as a Service (SaaS)

SaaS is also known as "On-Demand Software". It is a software distribution model in which services are hosted by a cloud service provider. These services are available to end-users over the internet so, the end-users do not need to install any software on their devices to access these services.

Services provided by SaaS providers

1. Business Services - SaaS Provider provides various business services to start-up the business. The SaaS business services include ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), billing, and sales.
2. Document Management - SaaS document management is a software application offered by a third party (SaaS providers) to create, manage, and track electronic documents.
3. Social Networks - As we all know, social networking sites are used by the general public, so social networking service providers use SaaS for their convenience and handle the general public's information.

4. Mail Services - To handle the unpredictable number of users and load on e-mail services, many e-mail providers offering their services using SaaS.

PUBLIC CLOUD VS PRIVATE CLOUD

Cloud computing is a way of providing IT infrastructure to customer, it is not just a set of products to be implemented. For any service to be a cloud service, the following five criteria need to be fulfilled:

1. On demand self-service: Decision of starting and stopping of service depends on customers without direct interaction with providers.
2. Broad Network Access: Service must be available to any device using any network.
3. Resource Pooling: Provider create a pool of resources and dynamically allocate it to customers.

4. Rapid Elasticity: The services provided by provider must be easily expandable and quick.
5. Measured Services: Provider must measure the usage of service and charge it accordingly. Tracking of usage is also helpful in improving services.

Public Cloud:

Computing in which service provider makes all resources public over the internet. It is connected to the public Internet. Service provider serves resources such as virtual machines, applications, storage, etc to the general public over the internet. It may be free of cost or with minimal pay-per-usage. It is available for public display, Google uses the cloud to run some of its applications like google docs, google drive or YouTube, etc.

It is the most common way of implementing cloud computing. External cloud service provider owns, operates and delivers it over the public network.

It is best for the companies which need an infrastructure to accommodate large number of customers and working on projects which has diverse organisation i.e., research institution and NGO etc.

Private Cloud:

Computing in which service provider makes all resources public over the internet. It only supports connectivity over the private network. It has only authentic users and single-occupant architecture. Google back-end data of the applications like Google Drive, Google docs or YouTube, etc is not available to the public, these types of data and applications run on a private cloud.

The infrastructure and services are maintained and deployed over a private network; hardware and software are dedicated only to a private company i.e., members of the special entity.

It is best for the companies which need an infrastructure which has high performance, high security and privacy due to its best adaptability and flexibility.

Public Cloud	Private Cloud
• Cloud Computing infrastructure shared to public by service provider over the internet. It supports multiple customers i.e., enterprises.	• Cloud Computing infrastructure shared to private organisation by service provider over the internet. It supports one enterprise.
• Multi-Tenancy i.e., Data of many enterprises are stored in shared environment but are isolated. Data is shared as per rule, permission and security.	• Single Tenancy i.e., Data of single enterprise is stored.
• Cloud service provider provides all the possible services and hardware as the user-base is world. Different people and organization may need different services and hardware. Services provided must be versatile.	• Specific hardware and hardware as per need of enterprise are available in private cloud.
• It is hosted at Service Provider site.	• It is hosted at Service Provider site or enterprise.
• It is connected to the public internet.	• It only supports connectivity over the private network.

COMPUTING ON DEMAND

On-demand computing is a delivery model in which computing resources are made available to the user as needed. The resources may be maintained within the user's enterprise, or made available by a cloud service provider. When the services are provided by a third-party, the term cloud computing is often used as a synonym for on-demand computing. The on-demand model was developed to overcome the common challenge to an enterprise of being able to meet fluctuating demands efficiently. Because an enterprise's demand on computing resources can vary drastically from one time to another, maintaining sufficient resources to meet peak requirements can be costly. Conversely, if an enterprise tried to cut costs by only maintaining minimal computing resources, it is likely there will not be sufficient resources to meet peak requirements.

The on-demand model provides an enterprise with the ability to scale computing resources up or down with the click of a button, an API call or a business rule. The model is characterized by three attributes: scalability, pay-per-use and self-service. Whether the resource is an application program that helps team members collaborate or additional storage for archiving images, the computing resources are elastic, metered and easy to obtain.

Many on-demand computing services in the cloud are so user-friendly that non-technical end users can easily acquire computing resources without any help from the organization's information technology (IT) department. This has advantages because it can improve business agility, but it also has disadvantages because shadow IT can pose security risks. For this reason, many IT departments carry out periodic cloud audits to identify greynet on-demand applications and other rogue IT.