ADC501

Cloud Computing

Brief History

- Idea of computing in cloud – traces back to the origins of utility computing – concept proposed by computer scientist John McCarthy in 1961

- 2002 – Amazon Web Services (AWS) – offer services that provide remotely provisioned storage, computing resources and business functionality

- Microsoft Azure

- Google Cloud

- hardware, runtime environment and resources for a user by paying money – ***PAY AS YOU GO***

- no upfront commitment requirement

Figure 1.2. Cloud computing vision.

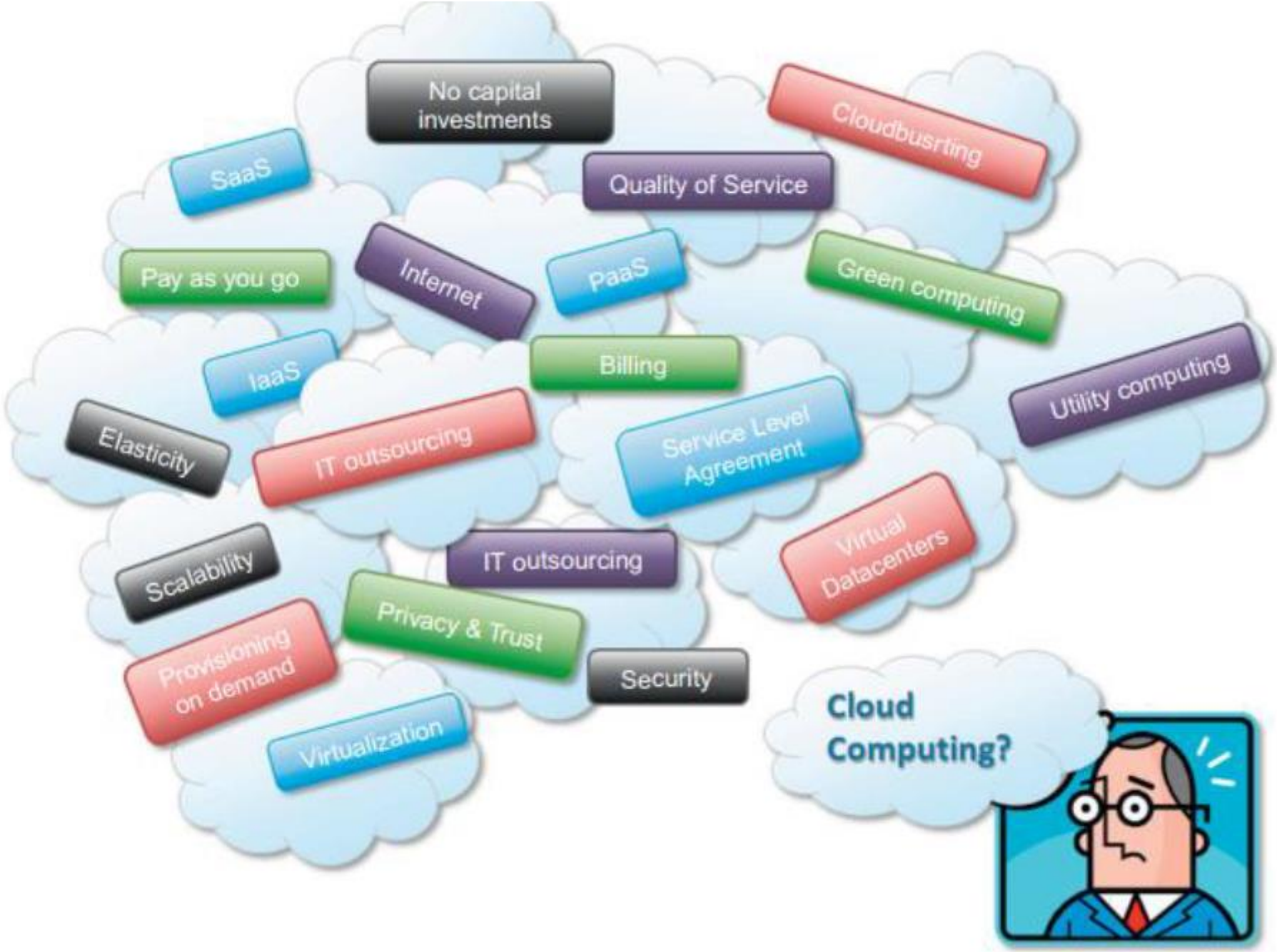


FIGURE 1.3 Cloud computing technologies, concepts, and ideas.

Definition of Cloud

- Gartner - " a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service to external customers using internet technologies"

- Forrestor Research – "a standardized IT capability (services, software or infrastructure) delivered via Interent technologies in a pay-per-use, self-service way"

- NIST – National Institute of Standards and Technology – sept 2011

  "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 cloud model is composed of five essential characteristics, three service models and four deployment models."

- 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

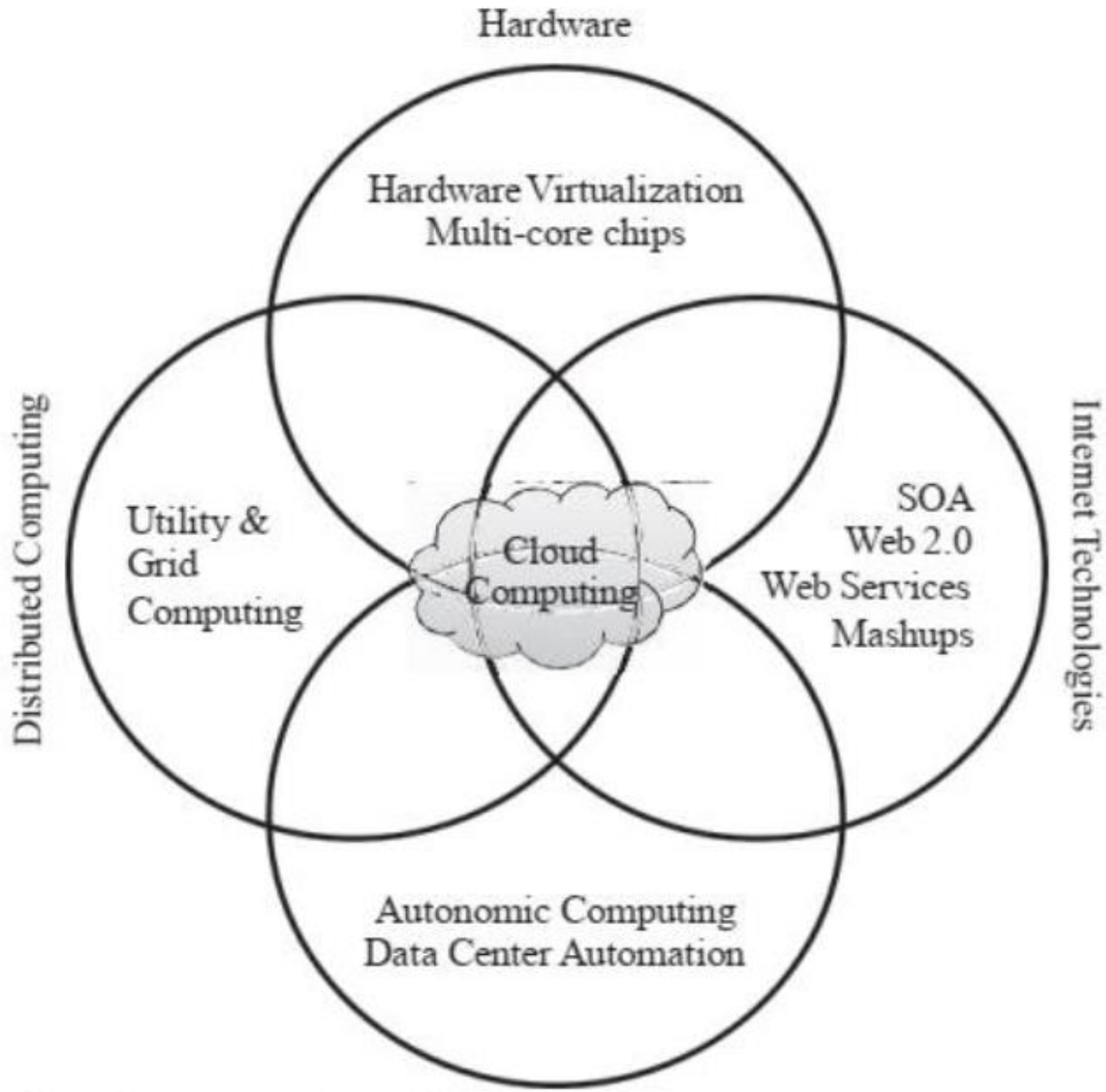


FIGURE 1.1. Convergence of various advances leading to the advent of cloud computing

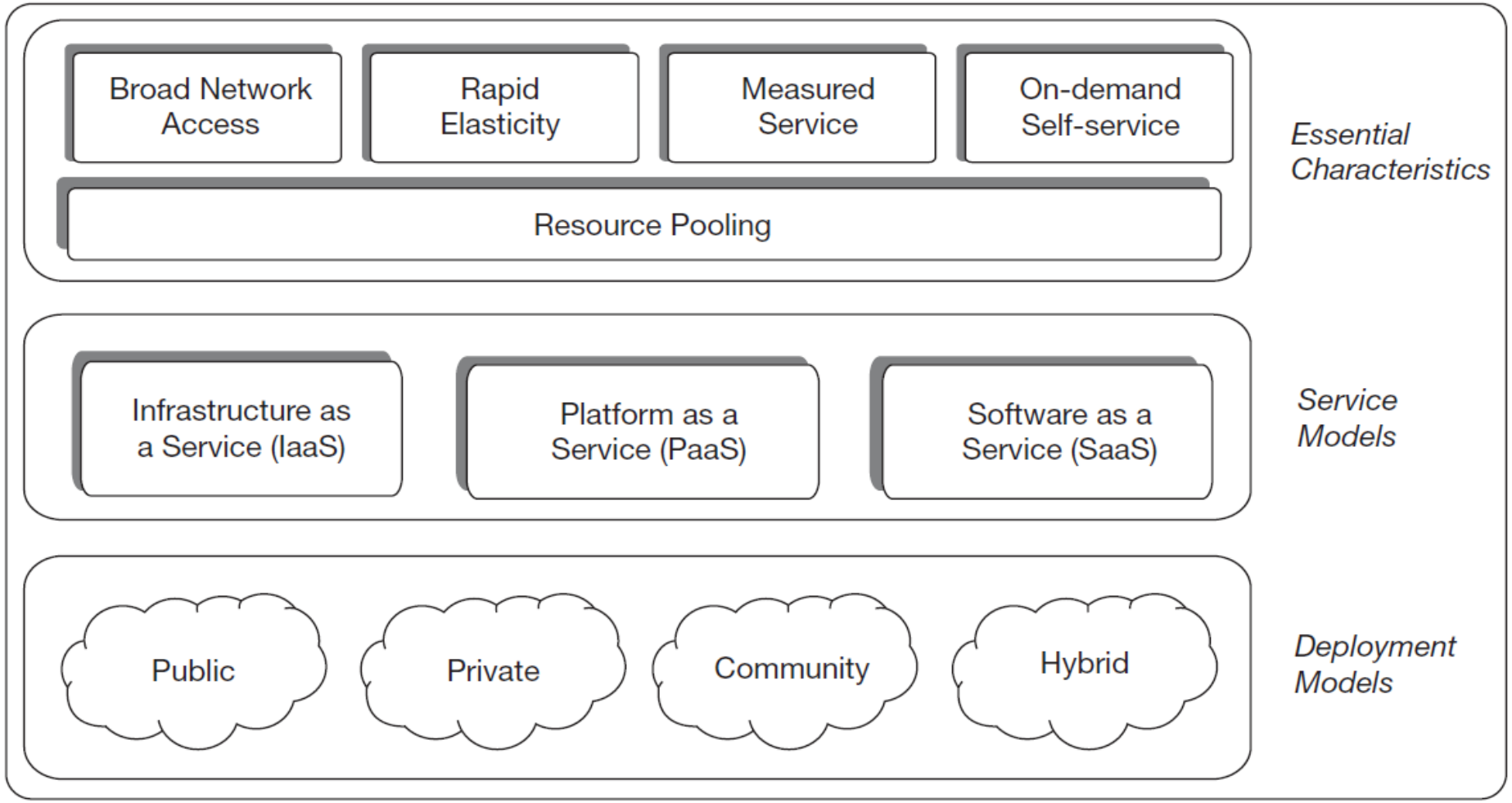


FIG 4.1: The NIST cloud computing model

NIST defines cloud computing by describing five essential characteristics, three cloud service models and four cloud deployment models.

Computing Paradigms

- **High Performance Computing**

a pool of processors/CPUS connected with other resources like memory, storage and input-output devices and the deployed software is enabled to run in the entire system of connected computers.
e.g. small group of computers to supercomputers,
Applications – solve scientific problems

- **Parallel Computing**

A problem is broken down into discrete parts that can be solved concurrently.
Instructions from each part are executed simultaneously on different processors.

- **Distributed Computing**

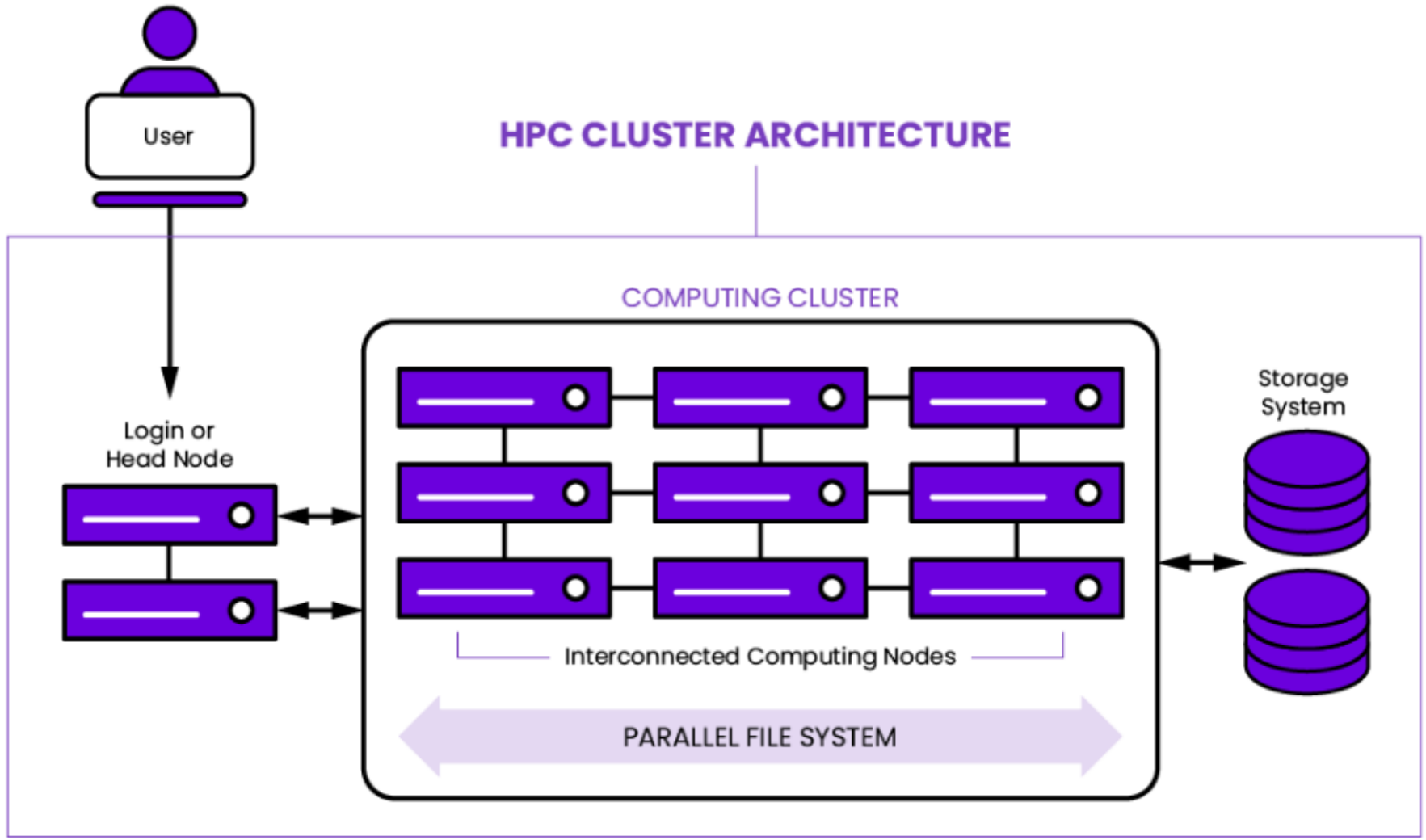
can be geographically distant and connected by a wide area network
Scalability , Redundancy or replication

- **Cluster Computing**

throughput, scaling speed, and processing
A cluster is a set of multiple computers that together operates as a single entity, called a node, to collectively work on the same set of tasks

- **Grid Computing**

computing resources in most of the organizations are underutilized but are necessary for certain operations , idea of grid computing is to make use of such nonutilized computing power by the needy organizations, and thereby the return on investment (ROI) on computing investments can be increased


HPC CLUSTER ARCHITECTURE
User
COMPUTING CLUSTER
Login or
Head Node
Storage
System
Interconnected Computing Nodes
PARALLEL FILE SYSTEM

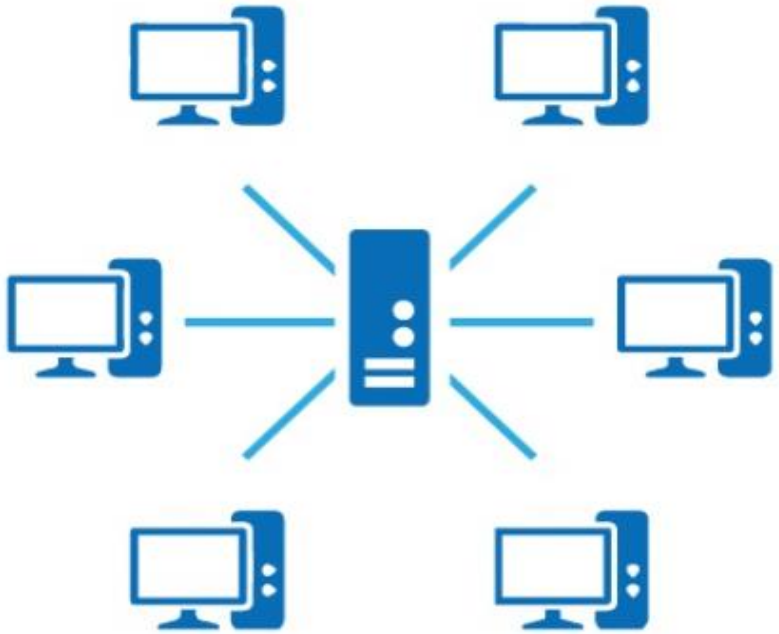
Autonomic Computing

- autonomic computing, which seeks to improve systems by decreasing human involvement in their operation
- IBM’s Autonomic Computing Initiative - self-configuration, self optimization, self-healing, and self-protection
- CC - management of service levels of running applications; management of data center capacity; proactive disaster recovery; and automation of VM provisioning

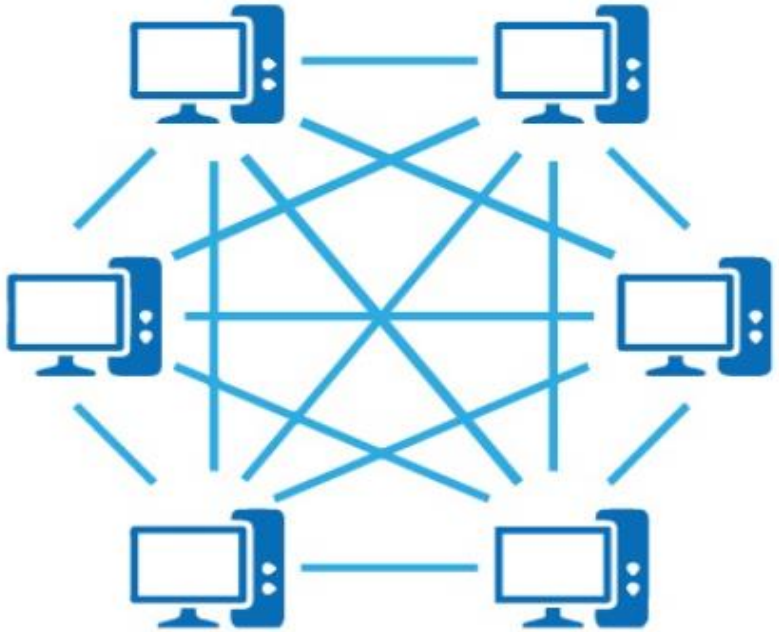
Other Models

- **Client – Server Model**

- **P2P – Peer to peer** decentralized network architecture in which participants (called peers) interact directly with each other without needing a central authority or server
  e.g. BitTorrent file-sharing network



A **Server** based Network



A **Peer-to-Peer** based Network

Virtualization

- Virtualization was first developed in the early 1960s
- used to partition large, **mainframe** hardware to improve utilization
- Large mainframes could be used to host up to 10,000 virtual machines
- enables an efficient use of resources and applications
- Does not provide load balancing across VMs nor does it improve communication between VMs

Hardware Virtualization

- running multiple operating systems and software stacks on a single physical platform

- a software layer, the virtual machine monitor (VMM), also called a hypervisor, mediates access to the physical hardware presenting to each guest operating system a virtual machine (VM)

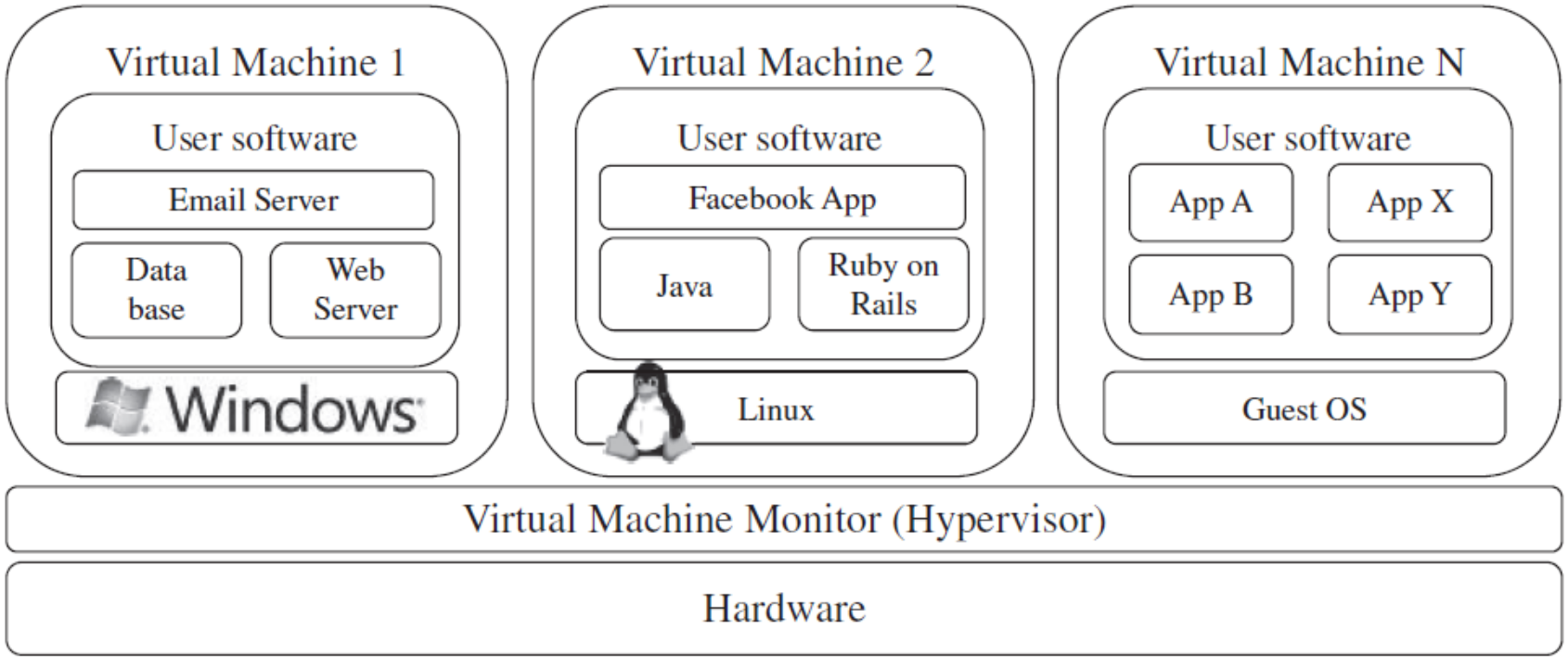


FIGURE 1.2. A hardware virtualized server hosting three virtual machines, each one running distinct operating system and user level software stack.

Cloud Deployment Models

Three deployment models

- **Public**

  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
  free or on-demand

- **Private**

  one organization's cloud service
  workloads deal with confidential information, intellectual property, and personally identifiable information (PII), medical records, financial data and other sensitive data.

- **Hybrid**

  contains links between a user's cloud (typically referred to as "*private cloud*") and a third-party cloud (typically referred to as "*public cloud*").

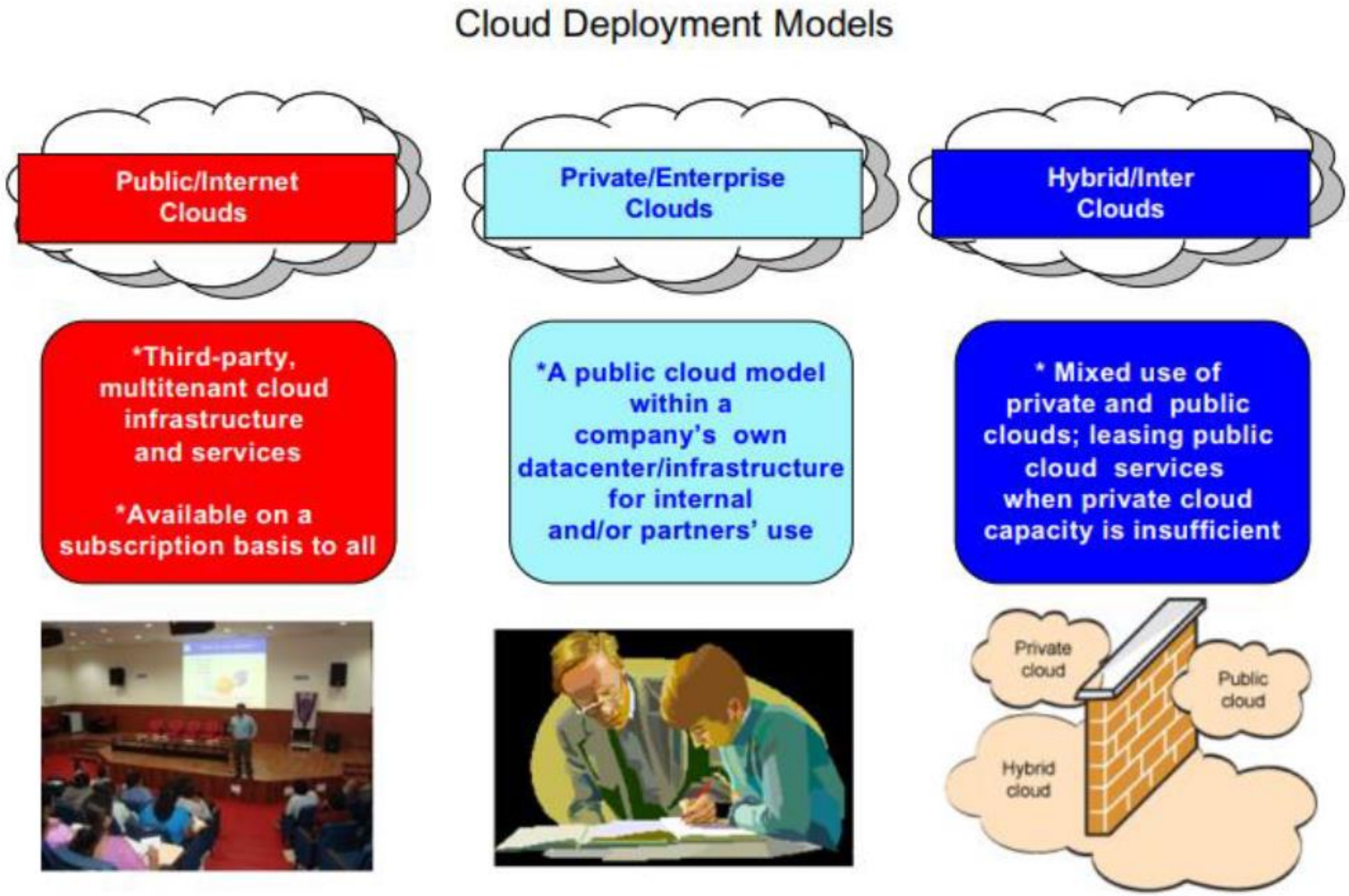


FIGURE 1.5 Major deployment models for cloud computing.

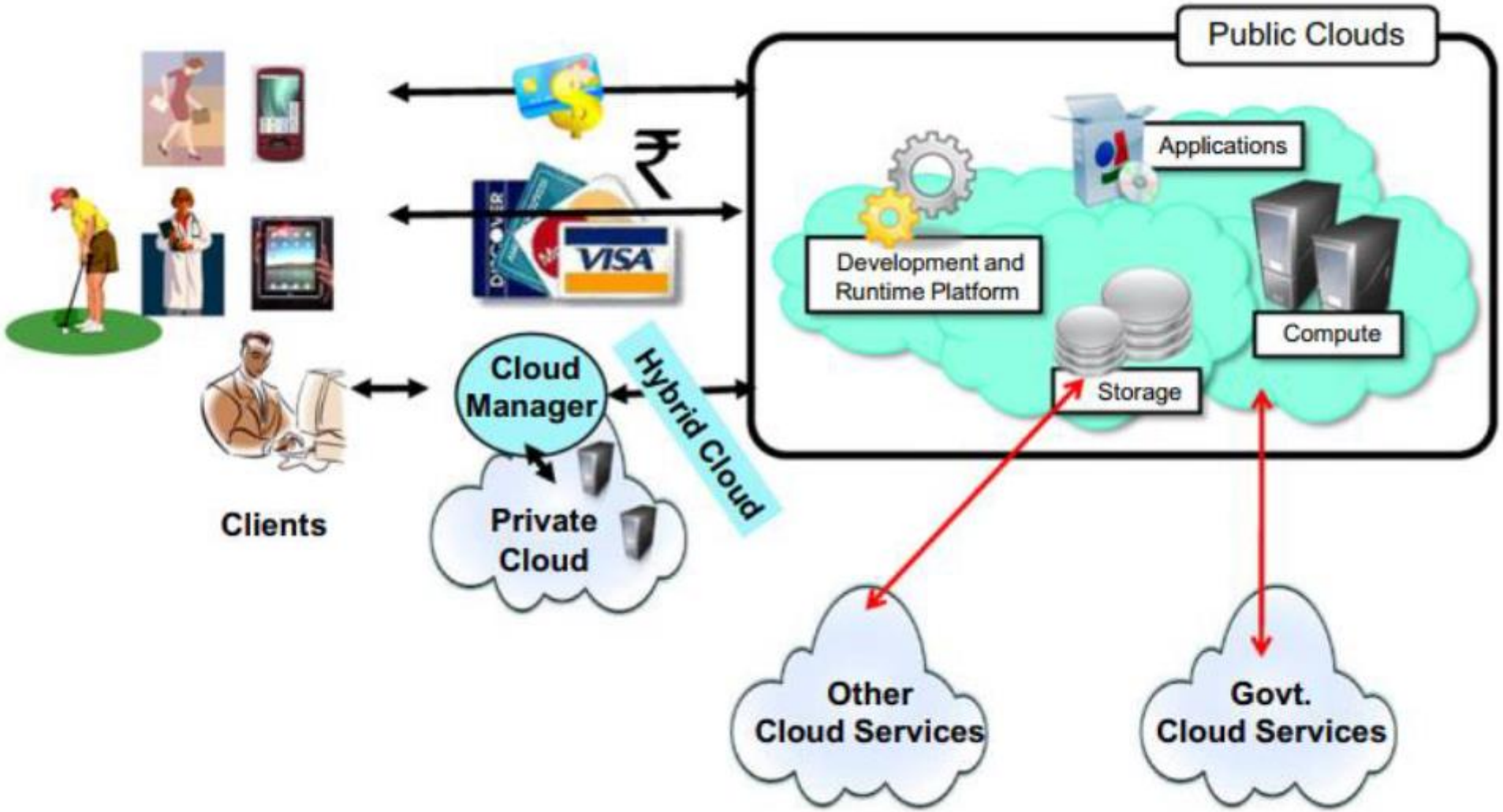


FIGURE 1.4 A bird's-eye view of cloud computing

Cloud Features

- 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

Business Drivers for Cloud Computing

- **Capacity Planning** - **capacity** – represents the maximum amount of work that an IT resource is capable of delivering in a given period of time.

  different strategies –

  Lead Strategy – adding capacity to an IT resource in anticipation of demand

  Lag Strategy – adding capacity when the IT resource reaches its full capacity

  Match Strategy – adding IT resource capacity in small increments, as demand increases

- **Cost Reduction** – cost of acquiring new infrastructure & cost of its ongoing ownership

  technical personnel for maintenance

  upgrades & patches of softwares

  power & cooling expenses

  security & access control measures

  administrative & accounts staff

- **Organizational Agility**

  measure of organization's responsiveness to change

  scale the IT resources as per the usage fluctuations

  IT resources to be more available and reliable

  keep up with market demands, competitive pressures and its own strategic business goals

Cloud Computing Services Classes

- Offering virtualized resources (computation, storage, and communication) on demand is known as **Infrastructure as a Service (IaaS)**
- A cloud platform offers an environment on which developers create and deploy applications and do not necessarily need to know how many processors or how much memory that applications will be using - **Platform as a Service (PaaS)**
  e.g. Google AppEngine
- Services provided by this layer can be accessed by end users through Web portals - **Software as a Service (SaaS)**

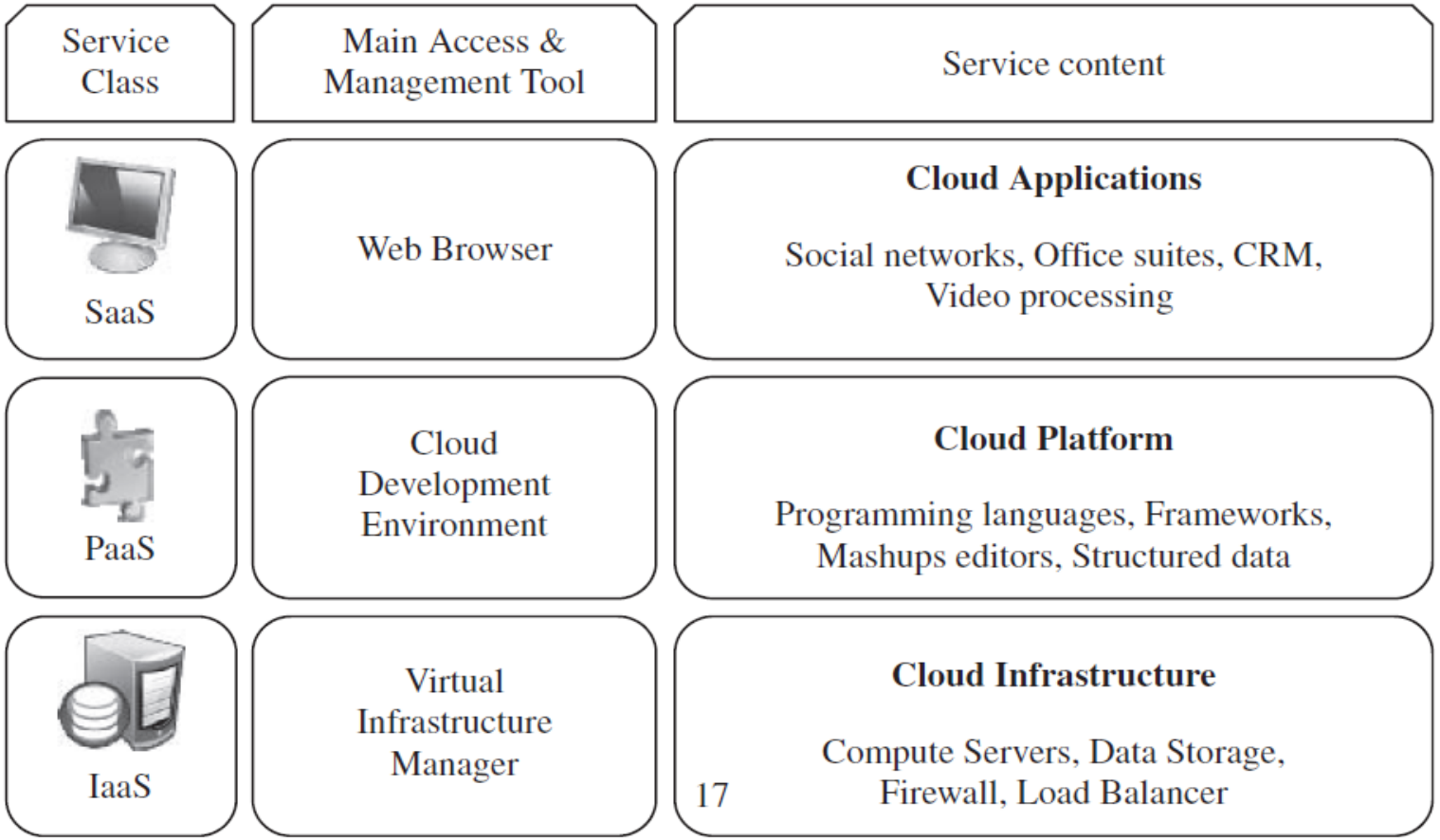


FIGURE 1.3. The cloud computing stack.

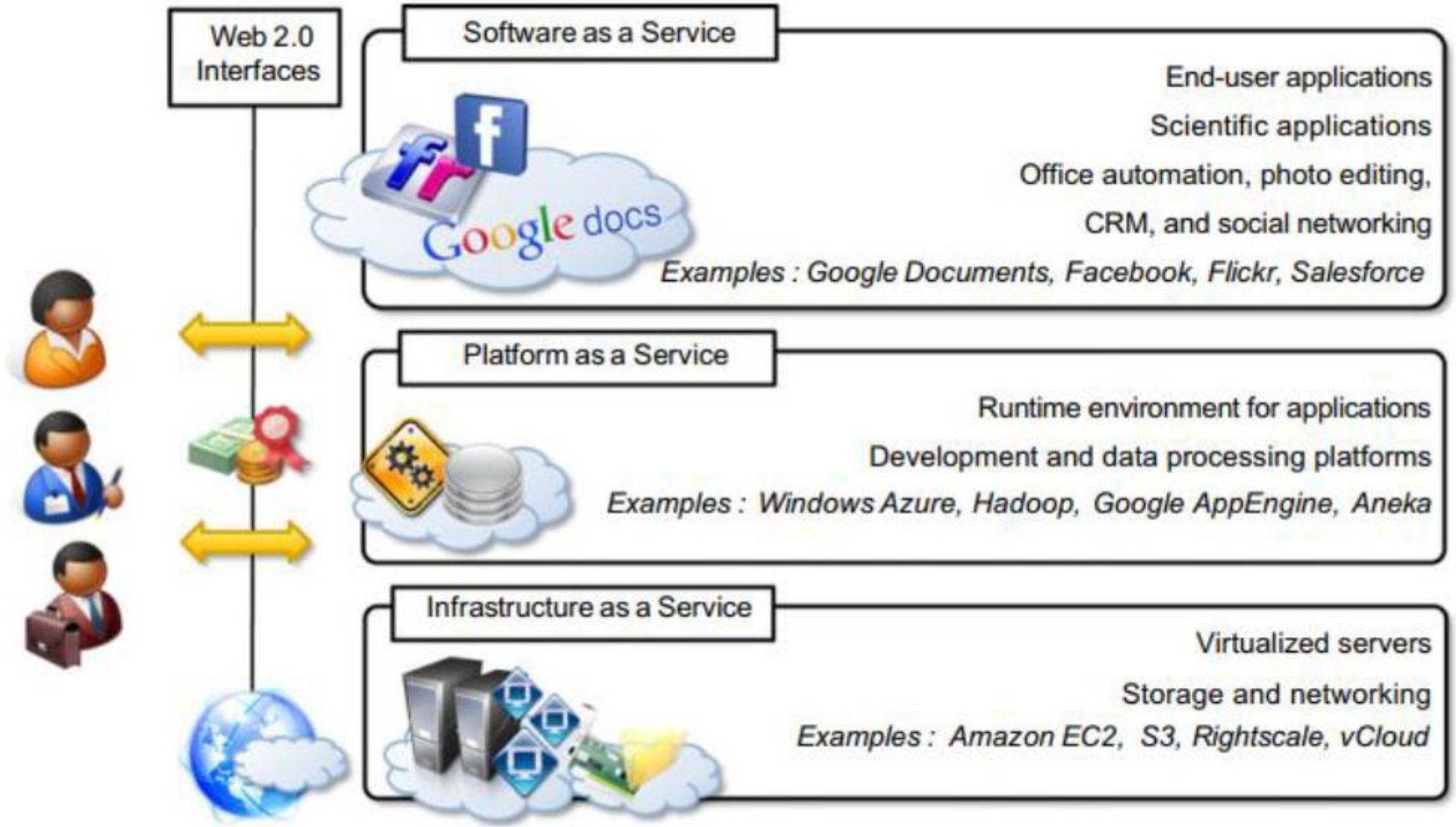


FIGURE 1.6 The Cloud Computing Reference Model.

Desired Features of Cloud

- **Self-Service**
- **Per-Usage Metering and Billing**
- **Elasticity**
- **Customization**
- **Performance monitoring & measuring**
- **Security**

Cloud API - **Cloud** APIs are software programs that transfer data between **cloud** computing services, or between **cloud** services and on-premise applications

Components of Cloud Computing

- Client - A client is an access device or software interface that a user can use to access cloud services

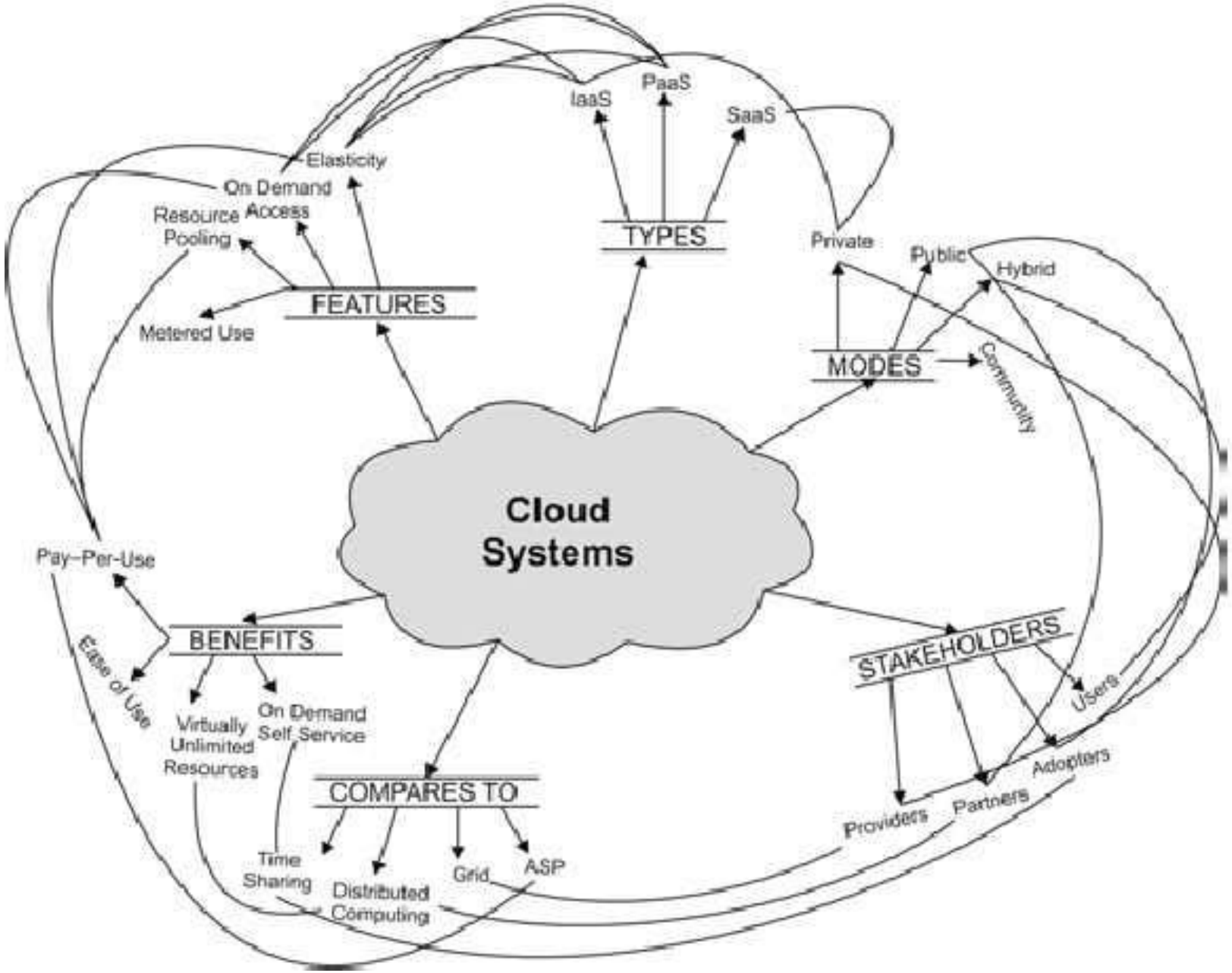
- Cloud network -
  A network is the connecting link between the user and cloud services

- Cloud Application programming Interfaces (APIs)

types of **cloud clients**

- Mobile clients
- Thin clients
- Thick clients

Various Aspects of Cloud Computing



When to avoid public cloud ?

The Cloud is not for Everyone

- Data privacy and confidentiality
- Legal liability and compliance
- Unpredictable availability of services
- Uneven performance and delayed response time to the end users
- Inability to use services when disconnected from the Internet