CS8791 – Cloud Computing

UNIT I INTRODUCTION

Introduction to Cloud Computing – Definition of Cloud – Evolution of Cloud Computing –Underlying Principles of Parallel and Distributed Computing – Cloud Characteristics – Elasticity in Cloud – On-demand Provisioning.

1.1 Introduction to Cloud Computing

1.1.1 What is Cloud Computing

- Cloud computing is the delivery of computing services over the internet. These services include:
    - Servers, Storage, Databases, Networking, Software, Analytics, Intelligence.
    - Cloud computing allows users to access these services on-demand. The services are hosted at a remote data center managed by a cloud services provider. Users can access the services from any computer with internet access.
- Cloud computing eliminates the need for users to manage physical resources themselves. Users only pay for what they use.

1.1.2 Top Cloud Service Providers

- Microsoft Azure
- Amazon Web Services (AWS)
- Google Cloud
- Alibaba Cloud
- IBM Cloud
- Oracle
- Salesforce
- SAP
- Rackspace Cloud
- VMWare

1.1.3 Advantages of Cloud Computing

- Cost Savings
- High Speed

- Back-up and restore data.
- Automatic Software Integration
- Reliability
- Mobility
- Unlimited storage capacity
- Collaboration

1.1.4 Disadvantages of Cloud Computing

- Security Threat in the Cloud
- Downtime
- Internet Connectivity
- Lower Bandwidth
- Vendor lock-in
- Limited Control

1.2 The Evolution of Cloud Computing

1.2.1 Hardware Evolution

- In 1930, binary arithmetic was developed.
  - computer processing technology, terminology, and programming languages
- In 1939, Electronic computer was developed.
  - Computations were performed using vacuum-tube technology.
- In 1941, Konrad Zuse's Z3 was developed.
  - Support both floating-point and binary arithmetic**.**

1.2.1.1 Generations of Computers

- First Generation Computers
- Second Generation Computers
- Third Generation Computers
- Fourth Generation Computers

1.2.1.1.1 First Generation Computers

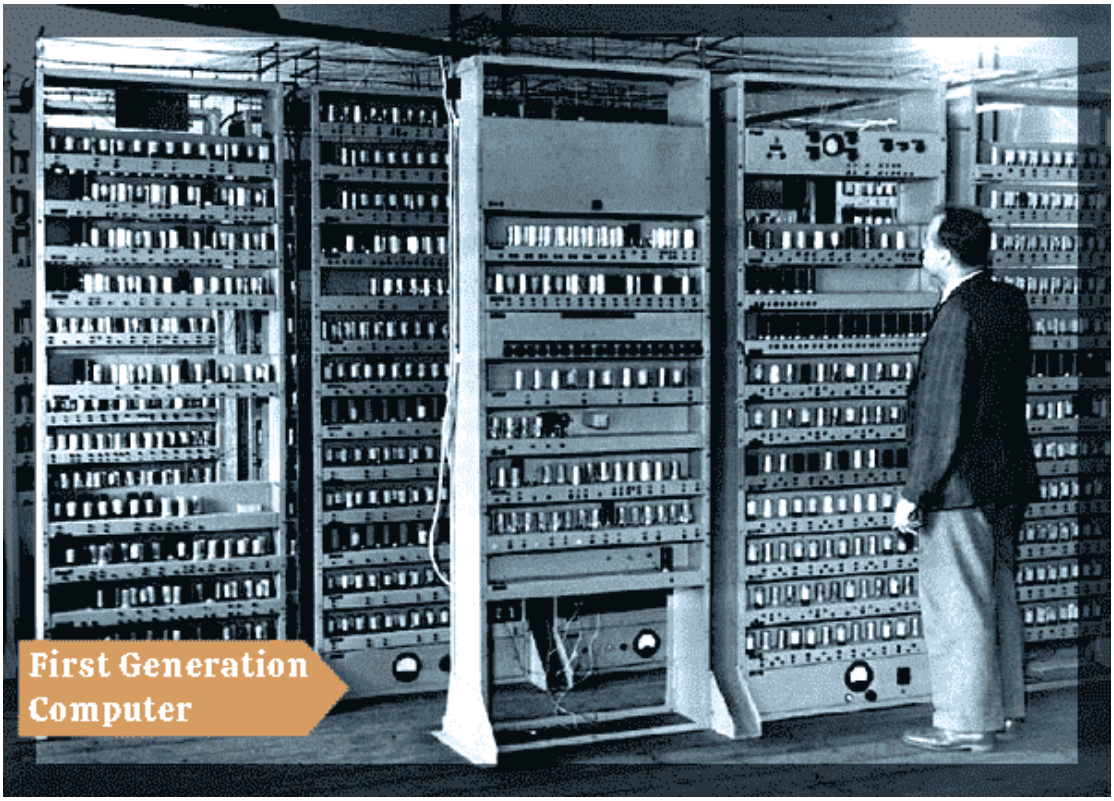
- Time Period : 1942 to 1955
- Technology: Vacuum Tubes
- Size : Very Large System
- Processing : Very Slow

- Examples:

1.ENIAC (Electronic Numerical Integrator and Computer)

2.EDVAC (Electronic Discrete Variable Automatic Computer)

- **Advantages:**
  - It made use of vacuum tubes which was the advanced technology at that time
  - Computations were performed in milliseconds.
- **Disadvantages:**
  - very big in size, weight was about 30 tones.
  - very costly.
  - Requires more power consumption.
  - A large amount of heat was generated.



1.2.1.1.2 Second Generation Computers

- Time Period : 1956 to 1965
- Technology: Transistors
- Size: Smaller
- Processing: Faster

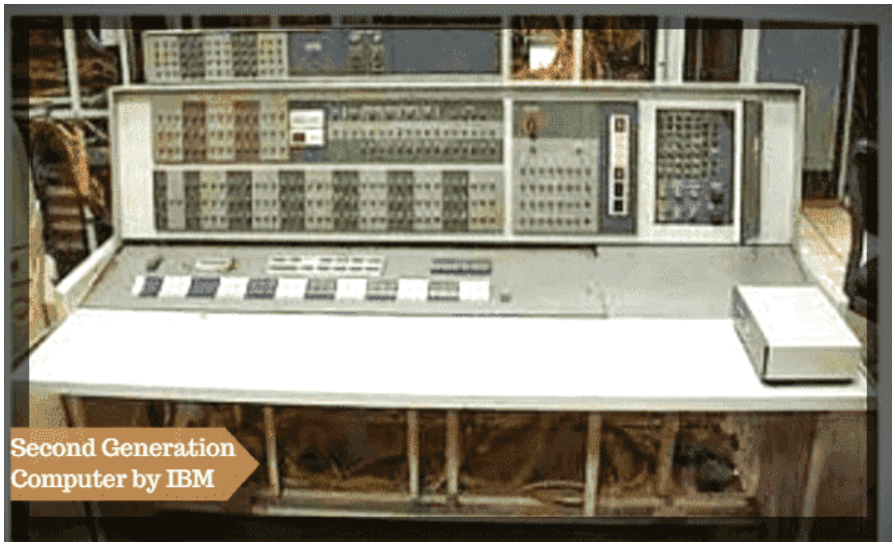
- Examples
  - Honeywell 400
  - IBM 7094

Advantages

- Less heat than the first generation.
- Assembly language and punch cards were used for input.
- Low cost than first generation computers.
- Computations were performed in microseconds.
- Better portability as compared to the first generation.

Disadvantages:

- A cooling system was required.
- Constant maintenance was required.
- Only used for specific purposes.



1.2.1.1.3 Third Generation Computers

- Time Period: 1966 to 1975
- Technology: ICs (Integrated Circuits)
- Size: Small as compared to 2nd generation computers
- Processing: Faster than 2nd generation computers

Examples

- ✓ PDP-8 (Programmed Data Processor)
- ✓ PDP-11

Advantages

- These computers were cheaper as compared to second-generation computers.
- They were fast and reliable.
- IC not only reduces the size of the computer, but it also improves the performance of the computer.
- Computations were performed in nanoseconds.

Disadvantages

- IC chips are difficult to maintain.
- The highly sophisticated technology required for the manufacturing of IC chips.
- Air conditioning is required.

Fig: Third Generation Computers

1.2.1.1.4 Fourth Generation Computers

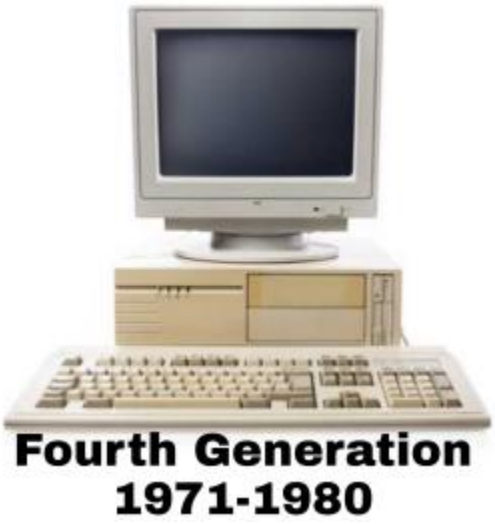
- Time Period : 1975 to Till Date
- Technology : Microprocessor
- Size : Small as compared to third generation computer
- Processing : Faster than third generation computer
- *Examples*
  - IBM 4341
  - DEC 10

Advantages:

- Fastest in computation and size get reduced as compared to the previous generation of computer.
- Heat generated is small.
- Less maintenance is required.

Disadvantages:

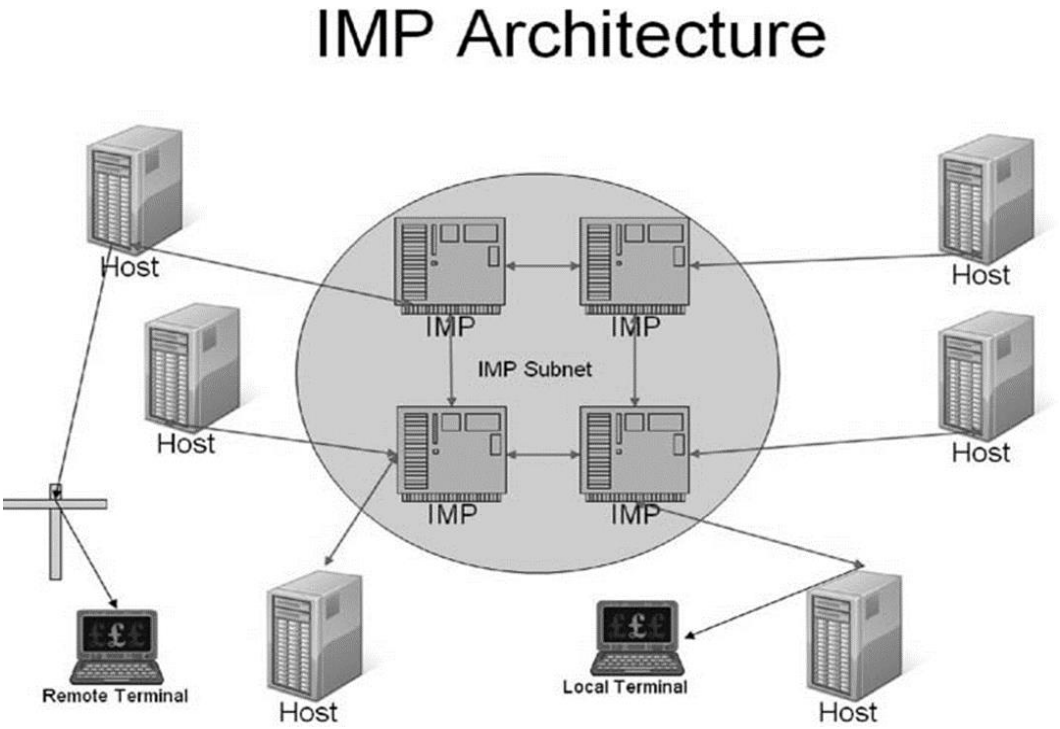
- The Microprocessor design and fabrication are very complex.
- Air conditioning is required in many cases due to the presence of ICs.
- Advance technology is required to make the ICs.



1.2.2 Internet Hardware Evolution

- Internet Protocol is the standard communications protocol used by every computer on the Internet.
- The conceptual foundation for creation of the Internet was significantly developed by three individuals.
  - Vannevar Bush – MEMIX (1930)
  - Norbert Wiener
  - Marshall McLuhan
- Licklider was founder for the creation of the ARPANET (Advanced Research Projects Agency Network)
- Clark deployed a minicomputer called an Interface Message Processor (IMP) at each site.

- Network Control Program (NCP)-first networking protocol that was used on the ARPANET.



1.2.2.1 Internet Hardware Evolution

- Establishing a Common Protocol for the Internet
- Evolution of Ipv6
- Finding a Common Method to Communicate Using the Internet Protocol
- Building a Common Interface to the Internet
- The Appearance of Cloud Formations—From One Computer to a Grid of Many

1.2.2.1.1 Establishing A Common Protocol for The Internet

- NCP essentially provided a transport layer consisting of the ARPANET Host-to-Host Protocol (AHHP) and the Initial Connection Protocol (ICP).

- **Application protocols**
  - File Transfer Protocol (FTP), used for file transfers,
  - Simple Mail Transfer Protocol (SMTP), used for sending email.
- **Four versions of TCP/IP**
  - TCP v1
  - TCP v2
  - TCP v3 and IP v3,

- TCP v4 and IPv4
- Today, IPv4 is the standard protocol.

1.2.2.1.2 Evolution of Ipv6

- IPv4 was never designed to scale to global levels.
- To increase available address space, it had to process large data packets (i.e., more bits of data).
- To overcome these problems, Internet Engineering Task Force (IETF) developed IPv6, which was released in January 1995.
- Ipv6 is sometimes called the Next Generation Internet Protocol (IPNG) or TCP/IP v6.

1.2.2.1.3 Finding Common Method to Communicate Using the Internet Protocol

- In the 1960s, the word hypertext was created by Ted Nelson.
- In 1962, Engelbart's first project was Augment, and its purpose was to develop computer tools to augment human capabilities.
- He developed the mouse, Graphical user interface (GUI), and the first working hypertext system, named NLS (oN-Line System).
- NLS was designed to cross-reference research papers for sharing among geographically distributed researchers.
- In the 1980s, the Web was developed in Europe by Tim Berners-Lee and Robert Cailliau.

1.2.2.1.4 Building A Common Interface to The Internet

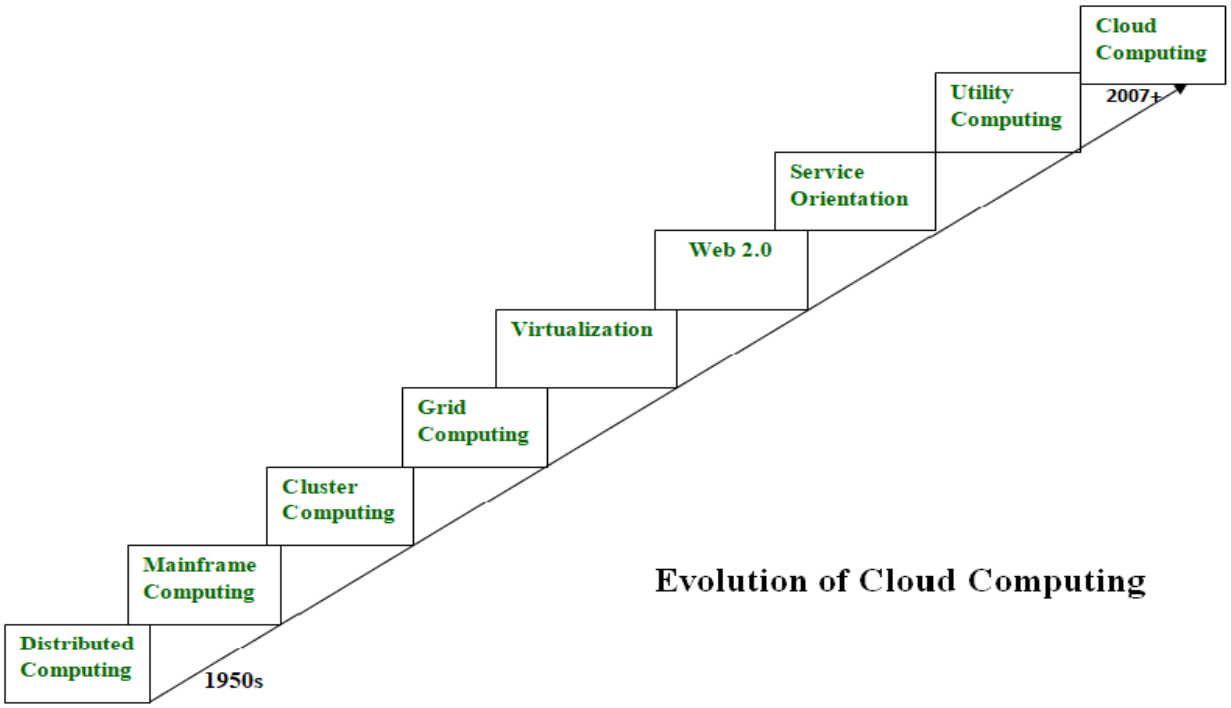
- Berners-Lee developed the first web browser featuring an integrated editor that could create hypertext documents.
- Following this initial success, Berners-Lee enhanced the server and browser by adding support for the FTP (File Transfer protocol)

Fig: First Web Browser

- Mosaic was the first widely popular web browser available to the general public. Mosaic support for graphics, sound, and video clips.
- In October 1994, Netscape released the first beta version of its browser, Mozilla 0.96b, over the Internet.
- In 1995, Microsoft Internet Explorer was developed that supports both a graphical Web browser and the name for a set of technologies.
- Mozilla Firefox, released in November 2004, became very popular almost immediately.

1.2.2.1.5 The Appearance of Cloud Formations -From One Computer to A Grid of many

- Two decades ago, computers were clustered together to form a single larger computer in order to simulate a supercomputer and greater processing power.
- In the early 1990s, Ian Foster and Carl Kesselman presented their concept of "The Grid." They used an analogy to the electricity grid, where users could plug in and use a (metered) utility service.
- A major problem in clustering model was data residency. Because of the distributed nature of a grid, computational nodes could be anywhere in the world.
- The Globus Toolkit is an open-source software toolkit used for building grid systems and applications.



Evolution of Cloud Computing

1.2.1.1.5 Evolution of cloud services

2008-2009	Google Application Engine Microsoft Azure
2006	S3 launches EC2
2002	Launch of Amazon Web Services
1990	The first milestone of cloud computing arrival of salesforce.com
1960	Super Computers Mainframes

1.2.1.1.6 Server Virtualization

- Virtualization is a method of running multiple independent virtual operating systems on a single physical computer.
- This approach maximizes the return on investment for the computer.
- Virtualization technology is a way of reducing the majority of hardware acquisition and maintenance costs, which can result in significant savings for any company.
    - Parallel Processing
    - Vector Processing
    - Symmetric Multiprocessing Systems
    - Massively Parallel Processing Systems

1.2.1.1.6.1 Parallel Processing

- Parallel processing is performed by the simultaneous execution of program instructions that have been allocated across multiple processors.
- Objective: running a program in less time
- The next advancement in parallel processing-multiprogramming
- In a multiprogramming system, multiple programs submitted by users, but each allowed to use the processor for a short time.
- This approach is known as “round-robin scheduling” (RR scheduling).

- RR Scheduling: All executable processes are held in a circular queue. The time slice is defined based on the number of executable processes that are in the queue.
- For example, if there are five user processes held in the queue and the time slice allocated for the queue to execute in total is 1 second then each user process is allocated 200 milliseconds of process execution time.

1.2.1.1.6.2 Vector Processing

- Vector processing was developed to increase processing performance by operating in a multitasking manner.
- Matrix operations were added to computers to perform arithmetic operations.
- This was valuable in certain types of applications in which data occurred in the form of vectors or matrices.
- In applications with less well-formed data, vector processing was less valuable.

1.2.1.1.6.3 Symmetric Multiprocessing Systems

- Symmetric multiprocessing systems (SMP) was developed to address the problem of resource management in master/slave models.
- In SMP systems, each processor is equally capable and responsible for managing the workflow as it passes through the system.
- The primary goal is to achieve sequential consistency.

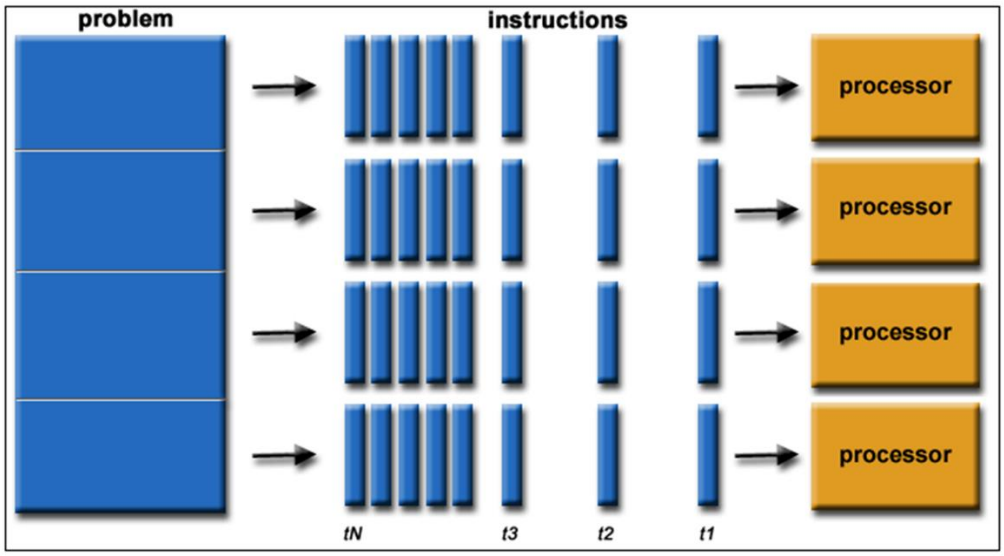
1.2.1.1.6.4 Massively Parallel Processing Systems

- In Massive parallel processing, a computer system with many independent arithmetic units or entire microprocessors, which run in parallel.
- All the processing elements are interconnected to act as one very large computer.
- Early examples of MPP systems were the Distributed Array Processor, the Goodyear MPP, the Connection Machine, and the Ultra computer.
- MPP machines are not easy to program, but for certain applications, such as data mining, they are the best solution.

1.3 Underlying Principles of Parallel and Distributed Computing

- Parallel computing refers to the process of breaking down larger problems into smaller, independent, often similar parts that can be executed simultaneously by multiple processors communicating via shared memory, the results of which are combined upon completion as part of an overall algorithm.
- The term parallel computing and distributed computing are often used interchangeably, even though they mean slightly different things.
- The term parallel implies a tightly coupled system, whereas distributed systems refers to a wider class of system, including those that are tightly coupled.
- More precisely, the term parallel computing refers to a model in which the computation is divided among several processors sharing the same memory.
- The architecture of parallel computing system is often characterized by the homogeneity of components: each processor is of the same type and it has the same capability as the others.
- The shared memory has a single address space, which is accessible to all the processors.
- Parallel programs are then broken down into several units of execution that can be allocated to different processors and can communicate with each other by means of shared memory.
- Originally parallel systems were considered as those architectures that featured multiple processors sharing the same physical memory and that were considered a single computer.
- Processing of multiple tasks simultaneously on multiple processors is called parallel processing.
- The parallel program consists of multiple active processes (tasks) simultaneously solving a given problem.
- A given task is divided into multiple subtasks using a divide-and-conquer technique, and each subtask is processed on a different central processing unit (CPU).
- Programming on multi-processor system using the divide-and-conquer technique is called parallel programming.
- Many applications today require more computing power than a traditional sequential computer can offer.

- Parallel Processing provides a cost effective solution to this problem by increasing the number of CPUs in a computer and by adding an efficient communication system between them.
- The workload can then be shared between different processors. This setup results in higher computing power and performance than a single processor a system offers.



1.3.1 Flynn‘s Taxonomy of Systems

SISD: Single instruction single data

SIMD: Single instruction multiple data

MISD: Multiple instructions single data

MIMD: Multiple instructions multiple data

1.3.1.1 SISD: Single Instruction Single Data

- An SISD computing system is a uniprocessor machine which is capable of executing a single instruction, operating on a single data stream.
- In SISD, machine instructions are processed in a sequential manner and computers adopting this model are popularly called sequential computers.
- All the instructions and data to be processed have to be stored in primary memory.
- Single Instruction: Only one instruction stream is being acted on by the CPU during any one clock cycle.
- Single Data: Only one data stream is being used as input during any one clock cycle

- This is the oldest type of computer.
- Examples: older generation mainframes, minicomputers, workstations and single processor/core PCs.

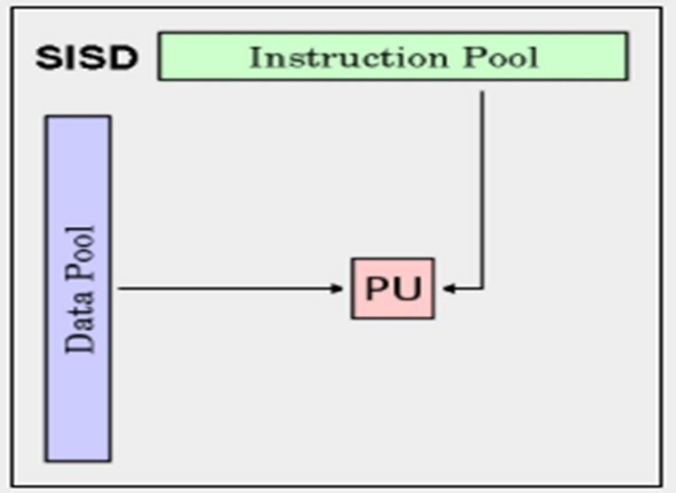


Fig: SISD: Single Instruction Single Data

1.3.1.2 SIMD: Single instruction multiple data

- A type of parallel computer.
- An SIMD system is a multiprocessor machine capable of executing the same instruction on all the CPUs but operating on different data streams.
- Single Instruction: All processing units execute the same instruction at any given clock cycle.
- Multiple Data: Each processing unit can operate on a different data element.
- Machines based on a SIMD model are well suited to scientific computing since they involve lots of vector and matrix operations.
- Most modern computers, particularly those with graphics processor units (GPUs) employ SIMD instructions and execution units.
- Processor Arrays: Thinking Machines CM-2, MasPar MP-1 & MP-2, ILLIAC IV
- Vector Pipelines: IBM 9000, Cray X-MP, Y-MP & C90, NEC SX-2, Hitachi S820, ETA10

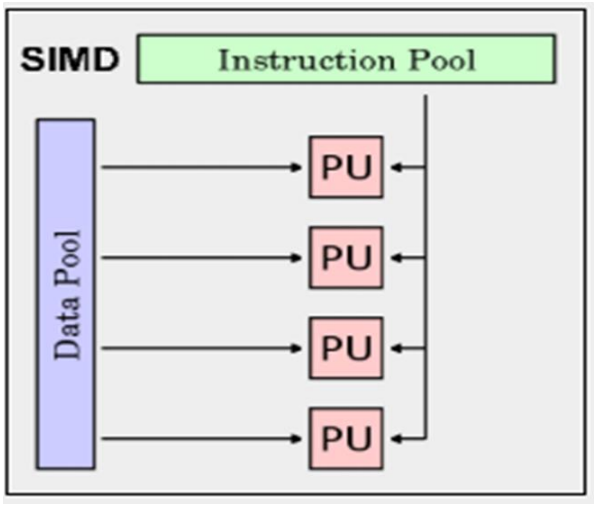


Fig: SIMD: Single Instruction Multiple Data

1.3.1.3 MISD: Multiple instructions single data

- An MISD computing system is a multiprocessor machine capable of executing different instructions on different PEs but all of them operating on the same dataset.
- The system performs different operations on the same data set.
- A type of parallel computer
- Multiple Instruction: Each processing unit operates on the data independently via separate instruction streams.
- Single Data: A single data stream is fed into multiple processing units.
- Machines built using the MISD model are not useful in most of the applications, a few machines are built, but none of them are available commercially.

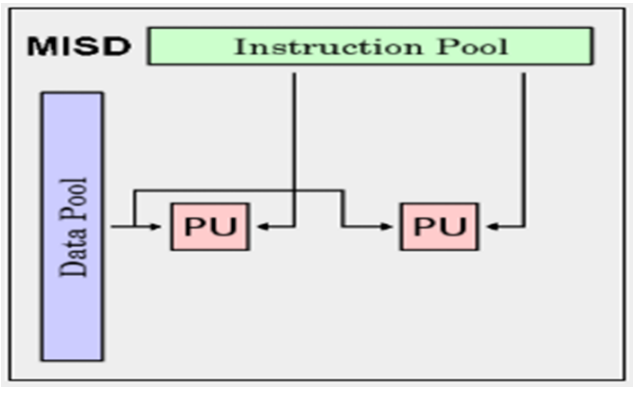


Fig: MISD: Multiple Instruction Single Data

1.3.1.4 MIMD: Multiple instructions multiple data

- An MIMD system is a multiprocessor machine which is capable of executing multiple instructions on multiple data sets.
- Multiple Instruction: Every processor may be executing a different instruction stream.
- Multiple Data: Every processor may be working with a different data stream.
- Examples: most current supercomputers, multi-core PCs.
- Example: IBM POWER5, Intel IA32

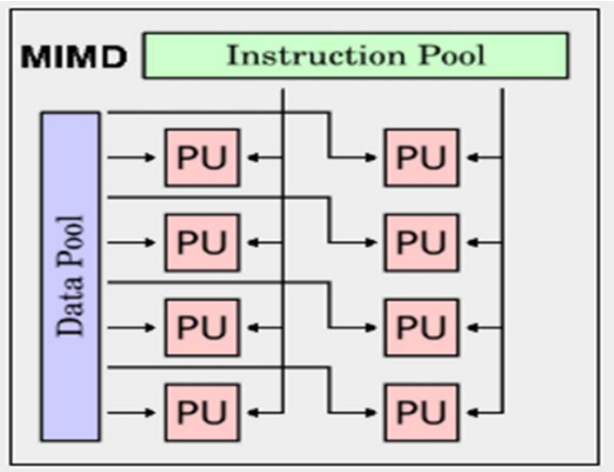
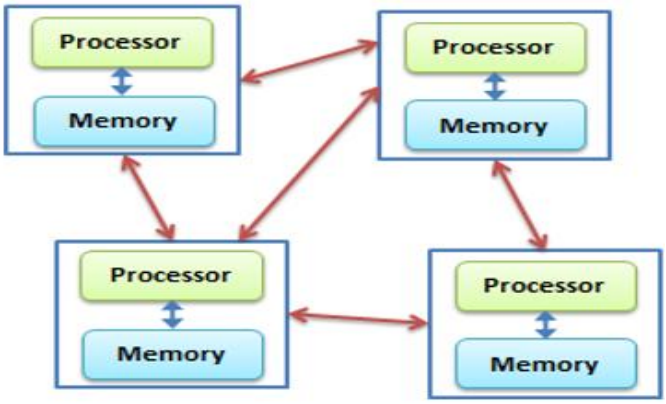


Fig: MISD: Multiple Instruction Multiple Data

1.3.2 Distributed Computing

1.3.2.1 What is Distributed Computing?

- Distributed computing is a model where multiple computers work together to solve a problem. In a distributed system, each computer has its own processing capabilities and may also store and manage its own data. The computers in a distributed system can be physically close together or geographically distant.



1.3.2.2 Elements of Distributed Computing

- Distributed computing is performed by the execution of multiple operations by multiple computers.

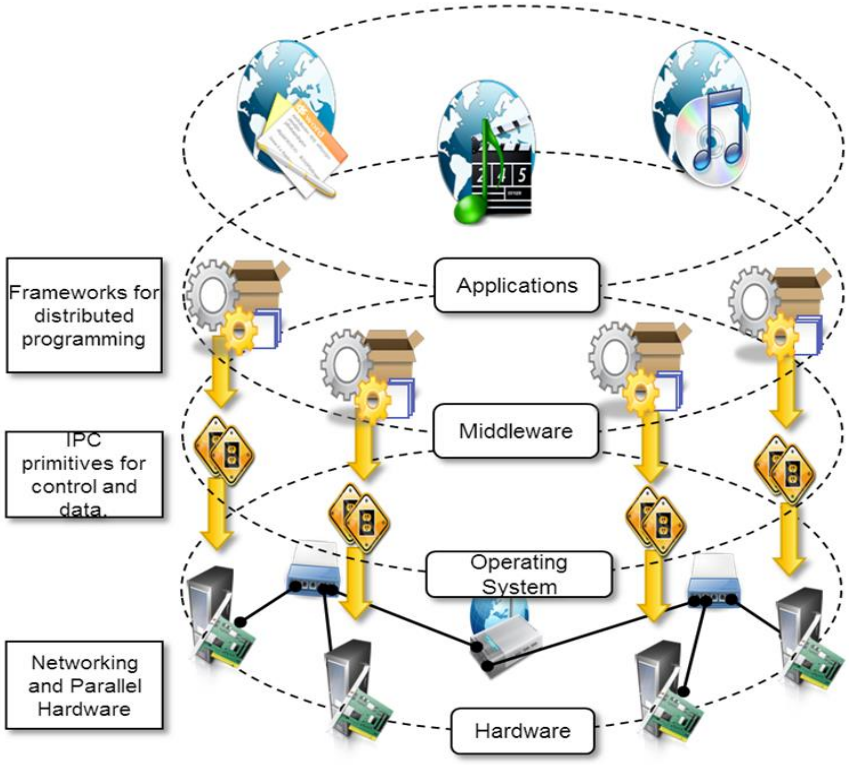
✓ 4 concepts

- General concepts and definitions
- Components of distributed system
- Architectural styles of distributed computing
- Models for inter-process communication

1.3.2.2.1 General concepts and definitions

- Distributed computing studies the models, architectures, and algorithms used for building and managing distributed systems.
- A distributed system is a collection of independent computers that communicate and coordinate their action only by passing messages.

1.3.2.2.2 Components of distributed system



- Layer1 (At bottom layer): Hardware
  - Physical infrastructure
- Layer2: Operating system
  - Process management and scheduling
- Layer3: Middleware
  - Provides transparency.
- Layer4 (at top layer): Applications

1.3.2.2.3 Architectural styles of distributed computing

- Middleware layer is the one that enables distributed computing, because it provides a coherent and uniform runtime environment.
- The use of well-known standards at the operating system level and hardware level allows easy harnessing of heterogeneous components and their organization into a coherent and uniform system.
- The architectural styles are classified into two major classes.
  - Software Architectural styles: Relates to the logical organization of the software.
  - System Architectural styles: Relates to the physical organization of distributed.

1.3.2.2.3.1 Software Architectural Styles

- Software architectural styles are based on the logical arrangement of software components.

Sl.No	Category	Most common Architectural Styles
1	Data Centered	Repository Blackboard
2	Data Flow	Pipe and filter Batch Sequential
3	Virtual Machine	Rule based Interpreter
4	Call and return	Main program and subroutine call/top-down systems Layered Systems
5	Independent Components	Communicating Processes Event Systems

1.3.2.2.3.2 Data Centered Architectures

- These architectures identify the data as the fundamental element of the software system
- Goal: integrity of data

Repository	Blackboard
The **repository architectural style** is the most relevant reference model in this category.	Blackboard models have become popular and widely used for artificial intelligent applications
It is characterized by two main components • central data structure • collection of independent components	• It is characterized by three main components: – Knowledge sources – Blackboard – Control
the dynamics of the system is controlled by independent components	These operate through a control shell that controls the problem-solving activity of the system

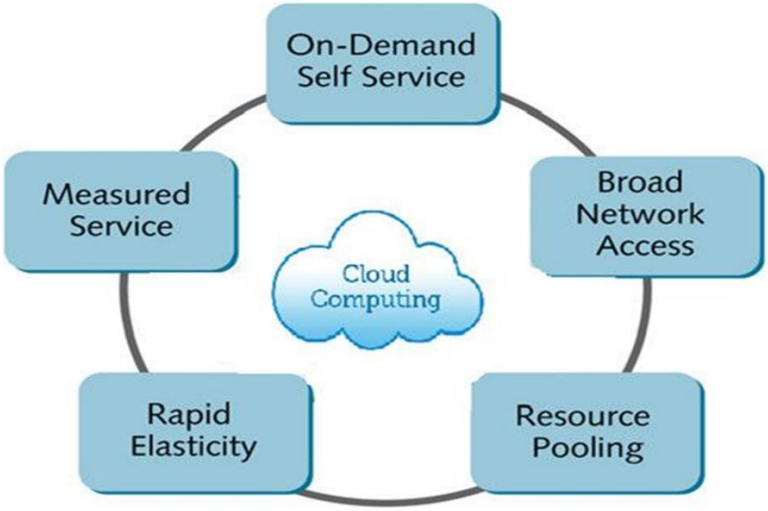
1.3.2.2.3.3 Data Flow Architectures

- These architectures identify the flow of data as the fundamental element of the software system.

Pipe-and-Filter	Batch Sequential
It has a independent filters for transformation and pipes for connection.	ordered sequence of separate programs executing one after the other.
Components are connected in a pipeline	input for the next program is the output generated by the last program
File grained	Coarse grained
Reduced latency due to the incremental processing of input	High latency
Localized input	External access to input
Concurrency possible	No concurrency
Interactivity	Non interactive

1.4 Cloud Characteristics

- The essential characteristics of the cloud computing model were defined by the National Institute of Standards and technology (NIST) and have since been redefined by a number of architects and experts.
- According to 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.

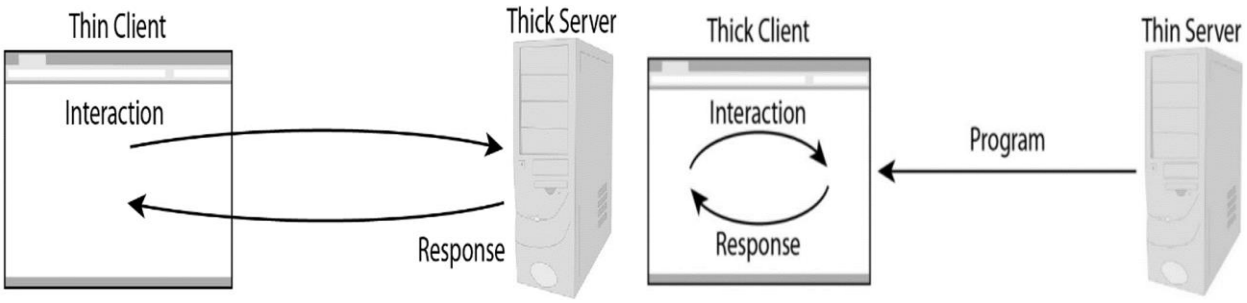


1.4.1 On-demand self-service

- 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.

1.4.2 Broad network access

- Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

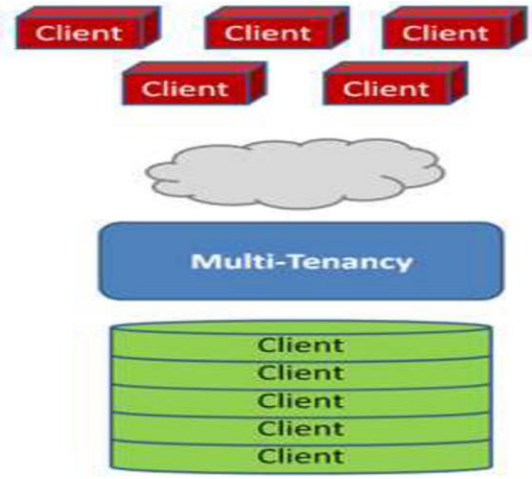


1.4.2.1 Key features Broad network access are

- High-speed network access.
- Uninterrupted availability of resources to any connected device.
- High performance of resources

1.4.3 Resource pooling

- The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model.



- 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.
- Examples of resources include storage, processing, memory, and network bandwidth.

1.4.4 Rapid elasticity

- Expand and shrink cloud computing resources at any time. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward according to demand.

- Unlimited availability of computing resources.

1.4.4.1 Elasticity has three major features.

- ✓ **Linear scaling**
  - The service can scale, independent of the number of users or workload size.
  - The performance experience for one of a thousand users is the same as for a single user.
- ✓ **On-demand utilization:**
  - user pays for consumption of the service based on the resource units consumed.
- ✓ **Pay-as-you-go:**
  - asset ownership is with the service provider, and the user pays for consumption of the service on the basis of the resource units consumed.

1.4.5 Measured service

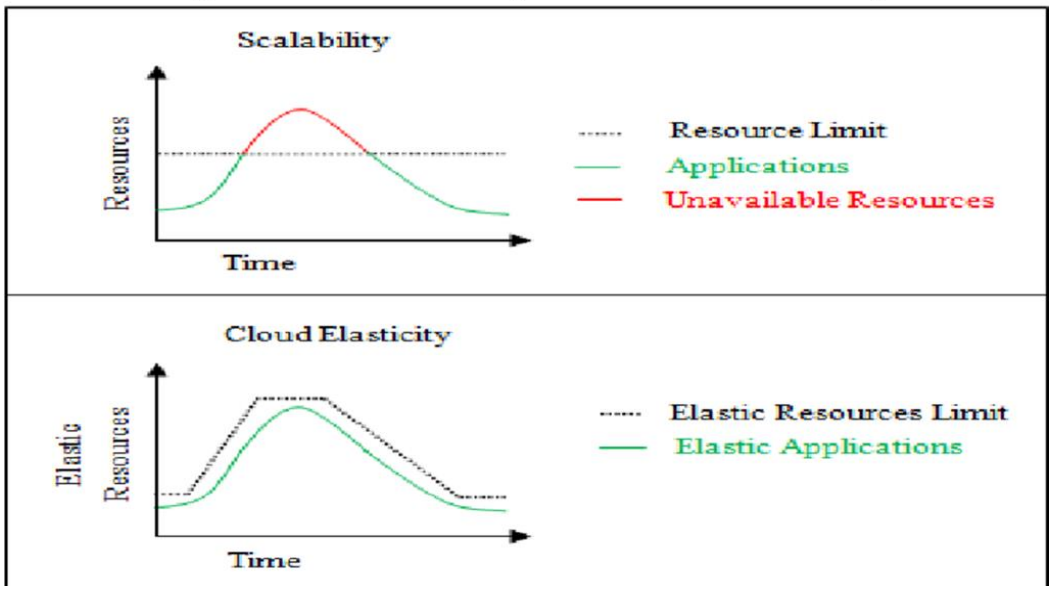
- ✓ Resource usage is monitored, measured, and reported (billed) transparently based on utilization. In short, you pay for use.
- ✓ Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.
- ✓ Any resource should be measurable.
- ✓ Dashboard to view the usage of resources by consumer.

1.5 Elasticity in cloud

"Ability to quickly scale in/out service."

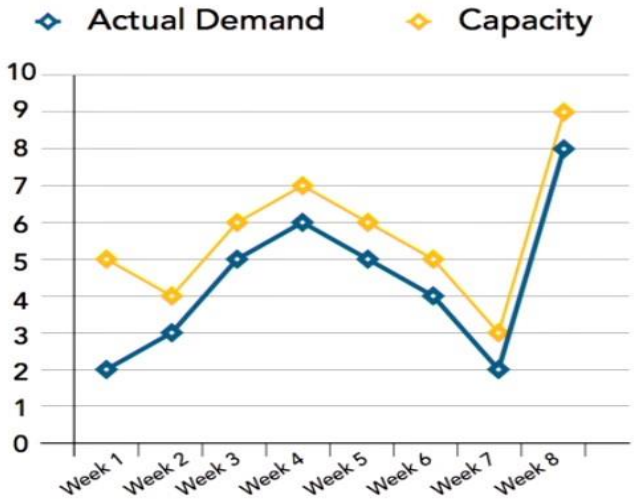
- ✓ Elasticity in cloud means the ability to quickly expand or decrease computer processing, memory and storage resources to meet changing demands without worrying about capacity planning and engineering for peak usage.

Figure 1: Scalability Vs Cloud Elasticity



- In scalability, there is a possibility of resources wastage and needed resources.
- But in Elasticity, there is no wastage of resources and needed resources.
- The elastic nature of the cloud refers to its ability to accommodate changes in load and demand of the system.

Cloud Elasticity



- Week1 : capacity– 5 servers, demand – 2 severs
- Week2 : capacity– 4 servers, demand – 3 severs
- Week3 : capacity– 6 servers, demand – 5 severs
- Week4 : capacity– 7 servers, demand – 6 severs
- Week5 : capacity– 6 servers, demand – 5 severs
- Week6 : capacity– 5 servers, demand – 4 severs
- Week7 : capacity– 3 servers, demand – 2 severs
- Week6 : capacity– 9 servers, demand – 8 severs

7

- With cloud elasticity, a company avoids paying for unused capacity or idle resources and does not have to worry about investing in the purchase or maintenance of additional resources and equipment.
- While security and limited control are concerns to take into account when considering elastic cloud computing, it has many benefits.
- Elastic computing is more efficient than your typical IT infrastructure, is typically automated so it does not have to rely on human administrators around the clock and offers continuous availability of services by avoiding unnecessary slowdowns or service interruptions.

1.5.1 Elastic Cloud: The Benefits

- ✓ On-demand computing
- ✓ Pay only for what you use.
- ✓ Failover and fault tolerance
- ✓ Ease of implementation

1.5.2 Problems in Elasticity

Resources provisioning time

- One potential problem is that elasticity takes time.
- A cloud virtual machine (VM) can be acquired at any time by the user, however, it may take up to several minutes for the acquired VM to be ready to use.
- The VM startup time is dependent on factors, such as image size, VM type, data center location, number of VMs, etc.
- **Monitoring elastic applications**
  - ✓ Elastic applications can allocate and deallocate resources (such as VMs) on demand for specific application components.
- **Elasticity requirements**
  - ✓ When deploying applications in cloud infrastructures (IaaS/PaaS), requirements of the stakeholder need to be considered in order to ensure proper elasticity behavior.
- **Multiple levels of control**
  - ✓ For multi-level control, control systems need to consider the impact lower-level control has upon higher level ones and vice versa (e.g., controlling virtual machines, web containers, or web services in the same time), as well as conflicts

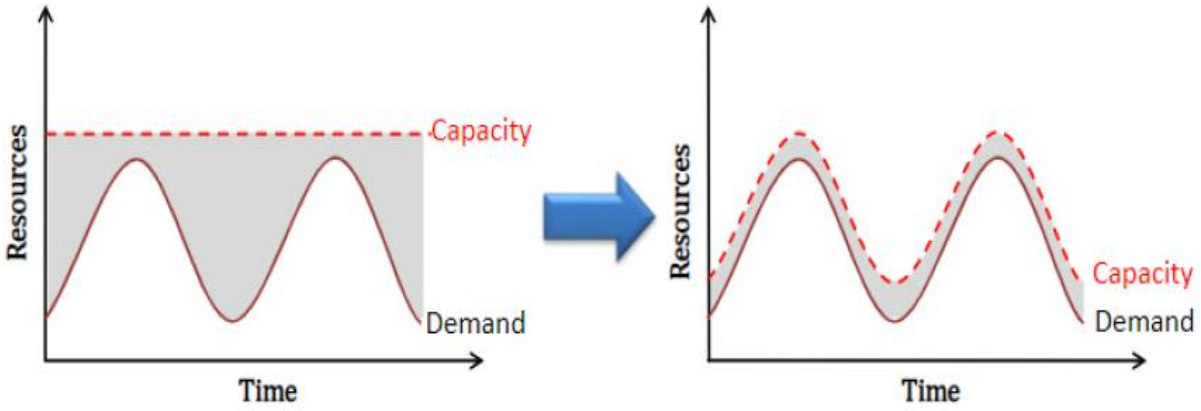
which may appear between various control strategies from various levels.Elasticity aims at matching the amount of resource allocated to a service with the amount of resource it actually requires, avoiding over- or under-provisioning.

- **Over-provisioning**
  - ✓ Allocating more resources than required should be avoided as the service provider often has to pay for the resources that are allocated to the service.
- **Under-provisioning**
  - ✓ Allocating fewer resources than required must be avoided, otherwise the service cannot serve its users with a good service. For example, under-provisioning the website may make it seem slow or unreachable.

1.6 On-demand computing

- On-demand computing is a delivery model in which computing resources are made available to the user as needed.
- 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.
- 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.
- Cloud computing services does not require any human administrators.
- For example-The consumer request is automatically processed by the cloud infrastructure, without human interaction.

- Cloud resources should be provisioned dynamically
  - Meet seasonal demand variations
  - Meet demand variations between different industries
  - Meet burst demand for some extraordinary events



- Week1 : capacity – 5 servers, demand – 2 severs
- Week2 : capacity – 4 servers, demand – 3 severs
- Week3 : capacity – 6 servers, demand – 5 severs
- Week4 : capacity – 7 servers, demand – 6 severs

- Week5 : capacity – 6 servers, demand – 5 severs
- Week6 : capacity – 5 servers, demand – 4 severs
- Week7 : capacity – 3 servers, demand – 2 severs
- Week6 : capacity – 9 servers, demand – 8 severs

1.6.1 Benefits of on-demand computing

- On-demand computing offers the following benefits:
- Flexibility to meet fluctuating demands. Users can quickly increase or decrease their computing resources as needed -- either short-term or long-term.
- Removes the need to purchase, maintain and upgrade hardware. The cloud service organization managing the on-demand services handles resources such as servers and hardware, system updates and maintenance.
- User friendly. Many on-demand computing services in the cloud are user friendly enabling most users to easily acquire additional computing resources without any help from their IT department. This can help to improve business agility.
- Cut costs. Saves money because organizations don't have to purchase hardware or software to meet peaks in demand. Organizations also don't have to worry about updating or maintaining those resources.