Cloud computing

Unit-1

⇒ systems modeling, clustering and virtualization:

* Scalable computing over the Internet:-
* The Age of Internet computing
* Scalable computing over the Internet.
* Technologies for network based systems.
* System models for Distributed and Cloud computing
* Performance
* Security and energy Efficiency.

⇒ LinPack Benchmarks are a measure of a System's Floating-Point computer Power

⇒ Two or more separating processing units are called Cores.

Scalable computing over the Internet:

what is scalability?

⇒ It is the ability of a computer application to continue to function when it is changed in size or volume.

Scalable computing over the Internet:

⇒ over the past 60 years, computing technology has undergone a series of platform and environment changes.

⇒ The evolutionary changes in ~~machine learning~~, machine Architecture, operating System platform network connectivity and Application workload

⇒ Billions of users use the Internet everyday.

⇒ Superconducto Supercomputer sites and large data centers must provide high performance computing services to huge numbers of Internet users concurrently.

⇒ upgrade data centers using fast server, storage systems, and high bandwidth networks.

⇒ using a centralized computer to solve computational problems, Parallel and distributed computers is used to solve large-scale problems over the Internet.

1. The Age of Internet Computing:-

⇒ Billions of people use the Internet every day.

⇒ As a result, supercomputer sites and large data centers must provide high-performance computing services to huge numbers of Internet users concurrently. (100, 1000)

⇒ Because of this high demand, the Linpack Benchmark (measure of a systems floating-point computing power) for High-Performance Computing (HPC) applications is no longer optimal for measuring system performance.

⇒ The emergence of computing clouds instead demands high-throughput computing (HTC) systems built with parallel and distributed computing technologies.

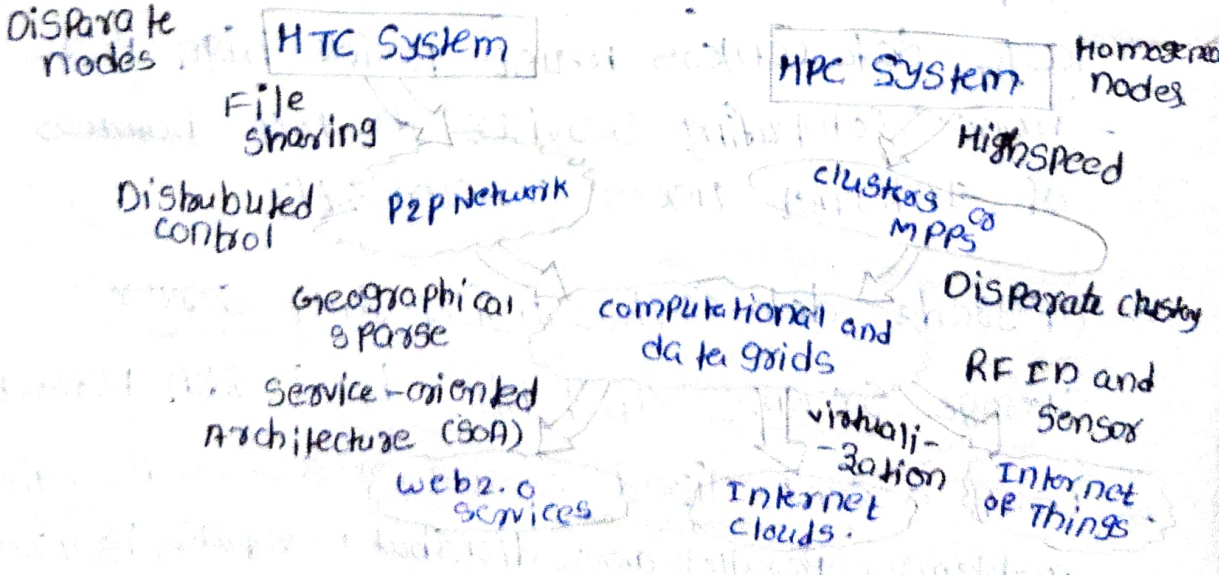


Fig: The evolution of HTC and HPC systems

1.2 High Performance computing

⇒ High Performance computing (HPC) is the ability to process data and perform complex calculations at high speed.

⇒ one of the best-known Types of HPC solutions is the super computer.

⇒ The speed of HPC systems has increased from Gflops (gigaflops) in the early 1990s to now Pflops in 2010 (peta flops)

[FLOPS :- Flotting Point Per second]

⇒ this improvement was driven mainly by the demands from scientific, engineering, and manufacturing communities.

1.3 High-Throughput computing:

⇒ High-throughput computing (HTC) is the use of distributed computing facilities for applications requiring large computing power over a long period of time.

⇒ The main aim of HTC to run a large number of computational tasks using resources in parallel.

⇒ HTC is mainly focus on increasing the overall throughput of the system by running many smaller size tasks parallelly.

⇒ HTC is commonly used in scientific research and engineering applications

Three New computing paradigms:

⇒ From service-oriented Architecture (SOA) provide web 2.0 service become available.

⇒ From virtualization make it possible to see the growth of Internet clouds.

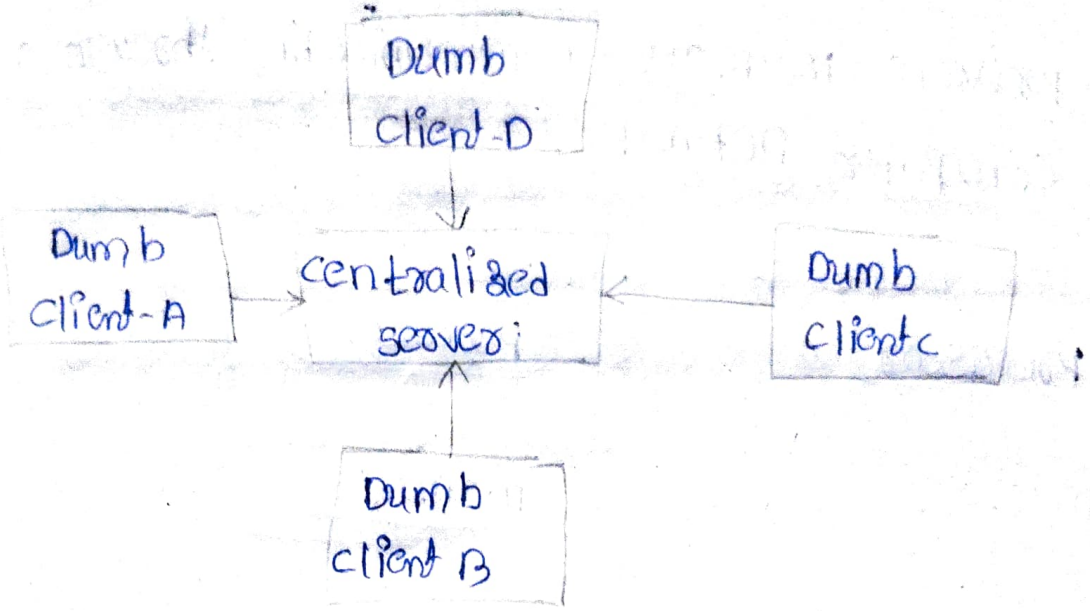
⇒ The Radio-Frequency identification (RFID) and sensor technologies provide Internet of Things.

Computing paradigm Distinctions:

1. Centralized computing:

⇒ This centralized computing paradigm all computer resources are centralized in one physical system

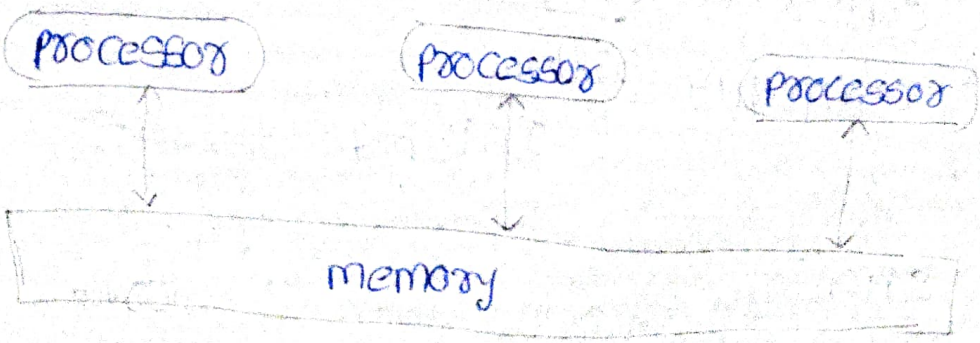
⇒ All resources processors, memory, storage) are fully shared and tightly coupled within one integrated os.



2. Parallel computing:

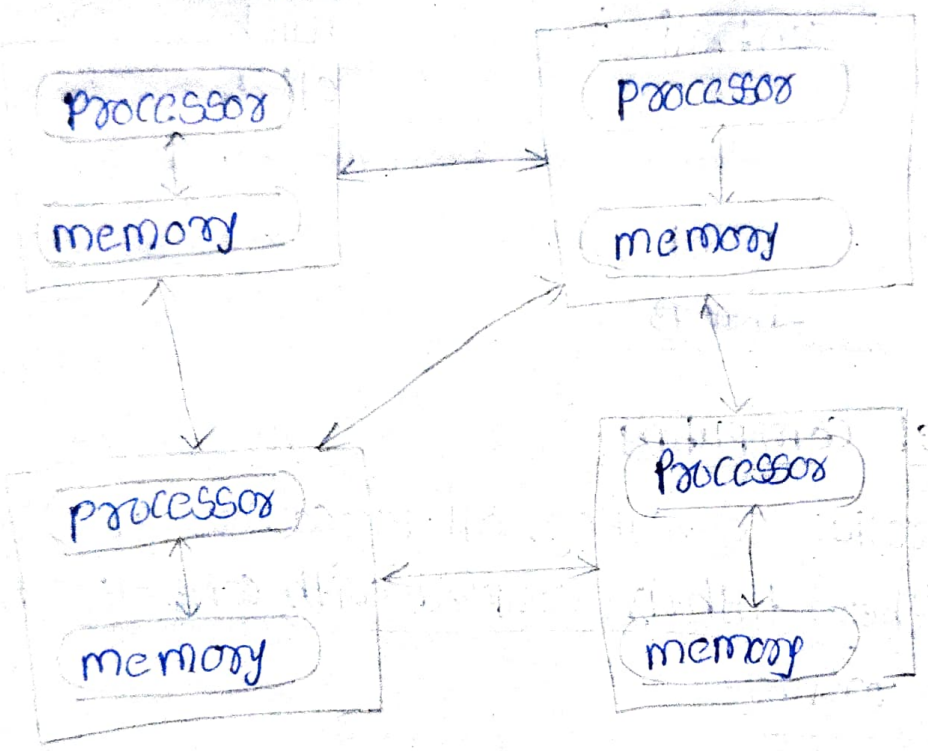
⇒ In parallel computing, all processors are either tightly coupled with centralized shared memory.

⇒ A computer system is capable to parallel computing is commonly known as a parallel computers.



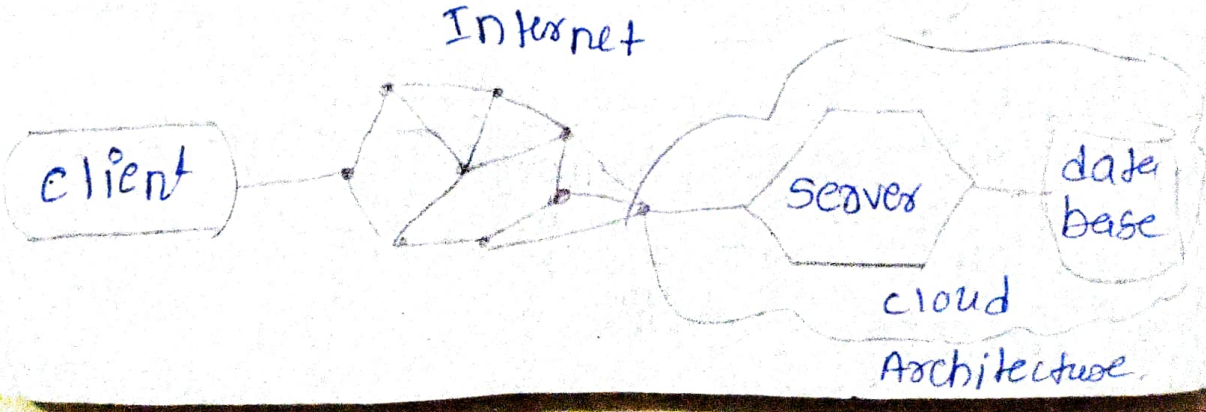
3. Distributed computing:-

⇒ A distributed system consists of multiple autonomous computers, each having its own private memory, communicating through a computer network.



4. Cloud computing:

⇒ An Internet cloud of resources can be either a centralized or a distributed computing system.



Architecture.

Degrees of parallelism:

⇒ Degrees of Parallelism is a measuring unit that defines the capability of a distributed system to run multiple programs in parallel.

Bit-level parallelism:- (BLP)

⇒ this type of systems, bit-level parallelism is used to transform bit-level processing into word level processing.

Instruction Level Parallelism:- (ILP)

⇒ When processing evolved from 4-bit to 64-bits
- ILP then came into Existence with which more than one instruction can be processed concurrently.

Data level Parallelism (DLP):-

⇒ DLP Depends on hardware and compiler support in order to carry out it's work efficiency

Task-Level Parallelism (TLP):-

⇒ It is not preferred over other typed of degrees of parallelism because it is complex in coding and compiling.

Job-Level Parallelism (JLP):-

⇒ Through the development process Degrees of parallelism is Transformed in terms of Job-Level Parallelism.

Technologies For Network-based Systems

1. Multicore CPUs and multithreading Technologies:-

multicore CPUs:-

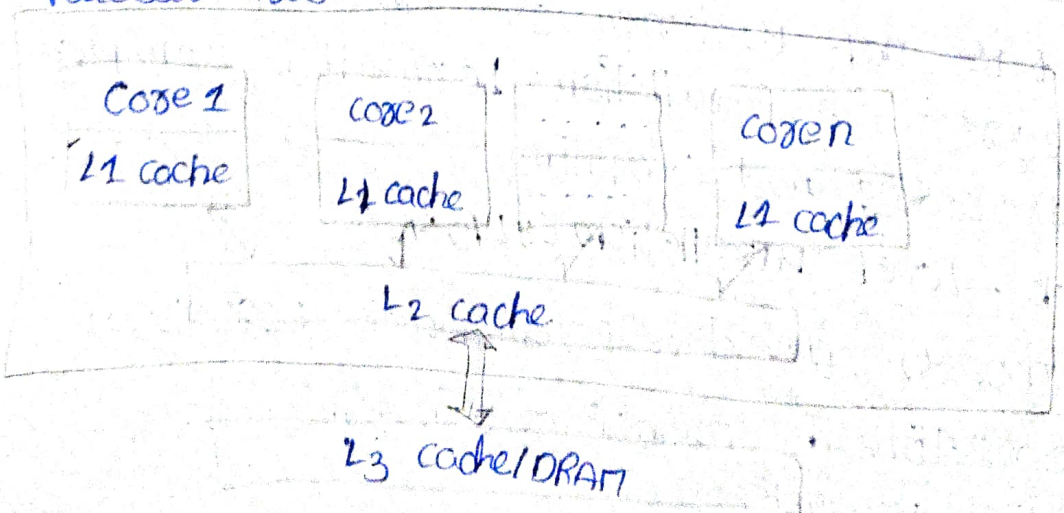
⇒ multicore CPUs make use of multiple cores within a single processor

⇒ In a multicore CPUs divide the processor into multiple cores such as dual, quad etc to carry out operations in parallel

⇒ The main advantage of this systems can be improve the potential performance of the overall system.

⇒ The examples of such systems in "Intel Processors" whose speed of processing increased from 10MHz to 4GHz.

⇒ Some systems use many-core GPUC Graphics Processing units) that make use of thousands of Processor cores.



* L1 cache is private, to individual cores
* L2 cache is common to all the cores

Multithreading:

⇒ multithreading is a feature which enables multiple threads to execute on a single processor in an overlapping manner.

⇒ A thread is an Atomic unit of a process and many threads usually make up a process.

⇒ In a multithreading environment, the resources of a processor are being shared by multiple threads.

⇒ Each thread has a separate Functional unit.

⇒ Functional unit including a separate Program counter, register file.

⇒ To enable multithreading, the hardware must be able to perform thread switching.

⇒ As thread are light weight they can execute and switch among themselves during the execution.

⇒ multithreading can be implemented in three ways.

1. Fine-grained multithreading:-

* The Threads are switched on each instruction
* The delay caused is very little.

* The subsequent thread is choosen using Round-robin fashion.
* In this method the thread is switched at every clock cycle.

2. Coars-Grained multithreading:-

* This method a thread only when a costly stall encountered.
* when a costly stall is executed without interrupt the any other thread.

3. Simultaneous multithreading :-

* this approach is implement on a super scalar multiprocessor.
* It is allows multiple threads to execute at the same time.

2. GPU computing, Exascale and Beyond:-

⇒ Gpu stands for "graphics processing unit" which is used to manipulate 3D graphics, multimedia and images.

⇒ The main aim of this process is free up the processor from processing

⇒ Implementing Gpu as a coprocessor on video card.

⇒ Gpu was first developed by NVIDAI in 1999.

⇒ It is processes multiple threads simultaneously

⇒ modern GPUs are capable of processing 512 concurrent threads.

⇒ GPU also used in processing floating-point operations and data-intensive calculation.

1. Fermi GPU:-

Fermi based GPU has the following Advantages:

* It has improved memory Access.
* It generate cache hierarchy
* It shares memory among streaming.

⇒ Fermi GPU consist the following components:

* 3.0 billion transistor.
* 512 cores arranged on 16 bit stream multiprocessor
* 384 bit DRAM interface is provided by GPU chip
* PCI express inorder to connect GPU to CPU.

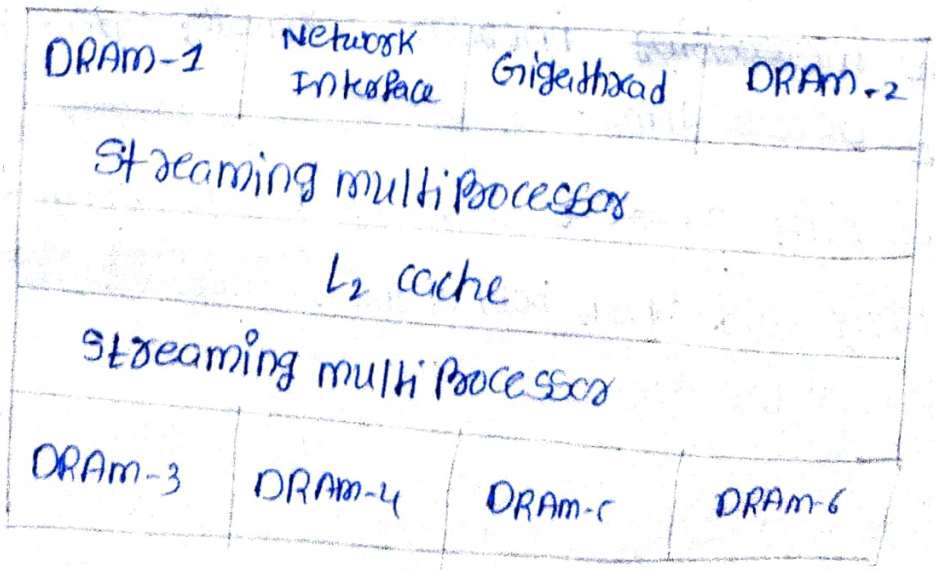
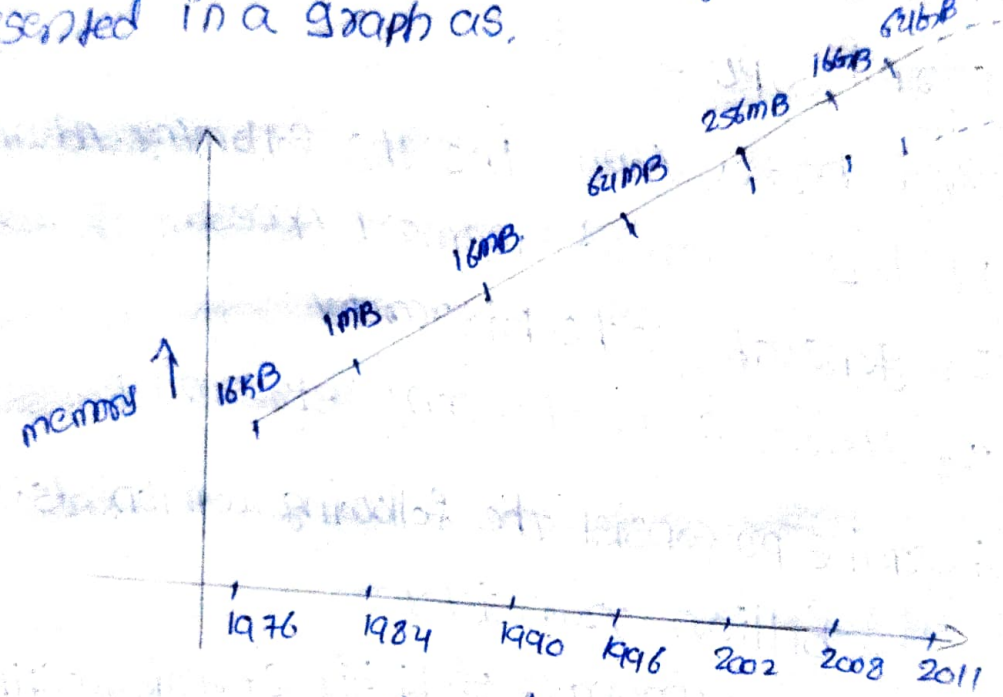


Fig: Fermi GPU

3. memory Technology, Storage Technology, Wide-Area networking:

memory Technology:-

⇒ The growth of memory Technology can be represented in a graph as,



⇒ From the above graph it can be observed that from 1976 to 2011, memory requirements increased from 16KB to 64GB.

⇒ ~~This increasing~~ increase Typically effect on the access time.

⇒ processing speed is directly Proportional to memory and when both of them increases in terms of memory.

Storage Technology:-

⇒ memory requirements for hard drive have increased from 260MB to 250GB during a 24 year period.

⇒ improve the performance of overall system by speeding up the applications associated with it.

Wide - Area Networking:-

⇒ Wide - Area Networks (WAN) spans a very large area that comprises of a country.

⇒ The ethernet speed has increased from 10mbps to 1Gbps.

⇒ Typically improves overall network performance as speed is getting 2x for every 1.5 years.

4. Virtual machines:-

⇒ A virtual machine is a representation of one or more computers on an existing physical computer.

⇒ It acquire's some space on the hard drive of the physical machine.

⇒ The virtual machine can perform all the task's of an operating system.

⇒ Low - level virtual machine operations

1. Multiplexing:-

operations carried out by various virtual machines.

2. suspension:-

It is possible to halt a virtual machine by suspending it temporarily and storage.

3. provision:

It provides flexibility and improves resource utilization.

Virtualization middle ware

⇒ virtualization middle ware runs between general-purpose hardware equipment and applications, providing virtualization functions for physical hardware.

Data Center Automation:

⇒ It is possible to Automate the system with respect to computing, networks and storage.

⇒ This typically makes the system effective and helps in generating quick result.

System models for distributed and cloud computing:

⇒ Distributed and cloud computing systems are built over a large number of Autonomous computer nodes.

⇒ these node machines are interconnected by SANs, LANs, WANs in a hierarchical manner.

⇒ today's networking technology, a few LAN switches can easily connected hundred of machines as a working cluster.

⇒ The WAN can connect many local clusters to form a very large cluster.

⇒ In this sense, one can build a massive system with millions of computers connected to edge network.

⇒ massive systems are considered highly scalable, and can reach web-scale connectivity, either physically or logically.

⇒ massive systems are classified into four groups: clusters, P2P networks, computing Grids, cloud platforms.

Functionality, Applications	computer clusters	Peer-to-Peer Networks	Data/computa-tional Grids	Cloud Platforms
Architecture, Network connectivity and size	Networks of compute nodes interconnected by LAN, WAN	Flexible network of client machine logically connected	Heterogeneous clusters interconnected by high-speed network.	visualized clusters of servers.
control and Resource management	Homogeneous nodes with distributed control	Autonomous client nodes	Centralized control	Dynamic Resources Provisioning

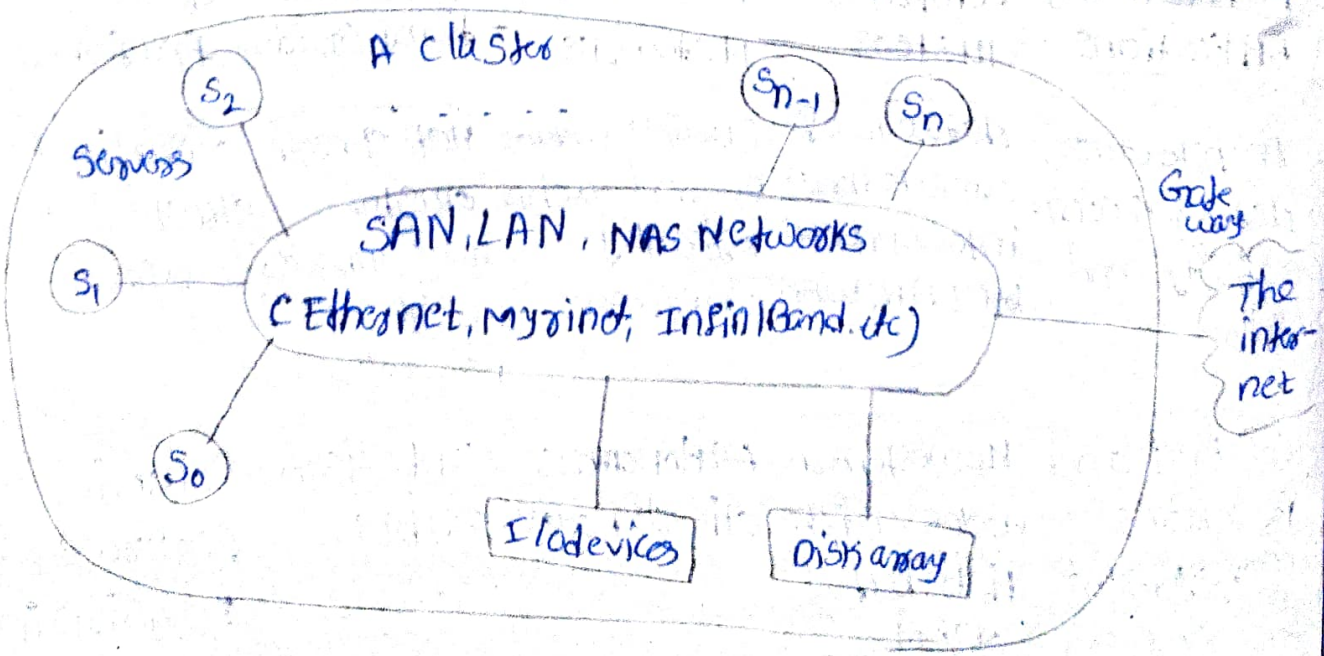
Applications and network centric service	High performance computing	most appealing to business file sharing	Distributed supercomputing	upgraded web search
Representative operational systems	Google search engine, IBM Road	Gnutella, eMule	Tera Grid, O-Grid	Google App engine, AWS

Fig: classification of paralled and distributed computing system.

Clusters of cooperative computers:

⇒ A computing cluster consists of interconnected stand-alone computers which work cooperatively as a single integrated computing resource.

Cluster Architecture:



⇒ This network can be simple as a SAN or a LAN to build a larger clusters with more nodes.

⇒ The clusters is connected to the internet via a virtual private network (VPN) gateway.

⇒ The gate way IP address locates the cluster.

Single System Image:-

⇒ Greg Pfister has indicated that an ideal cluster should merge multiple system images into a single system image (SSI).

⇒ This SSI support at various levels, including the sharing of CPUs, memory, and I/O across all cluster nodes.

Grid Computing Infrastructures:-

⇒ Internet services such as the Telnet command enables a local computer to connect to a remote computer.

⇒ Grid computing allow close interaction among Applications running on distant computers.

Computational Grids:-

⇒ Computing Grid offers an infrastructure that couples computers, middlewares special instruments and people and sensors together.

⇒ The grid is often constructed across LAN, WAN, or Internet backbone networks at a regional.

⇒ The computers used in a grid are primary work stations, servers, clusters, and supercomputers.

Grid Families:-

⇒ Grid Technology demands new distributed computing models middleware, support network protocols, hardware infrastructure It is followed by development by IBM, microsoft, HP, Dell.

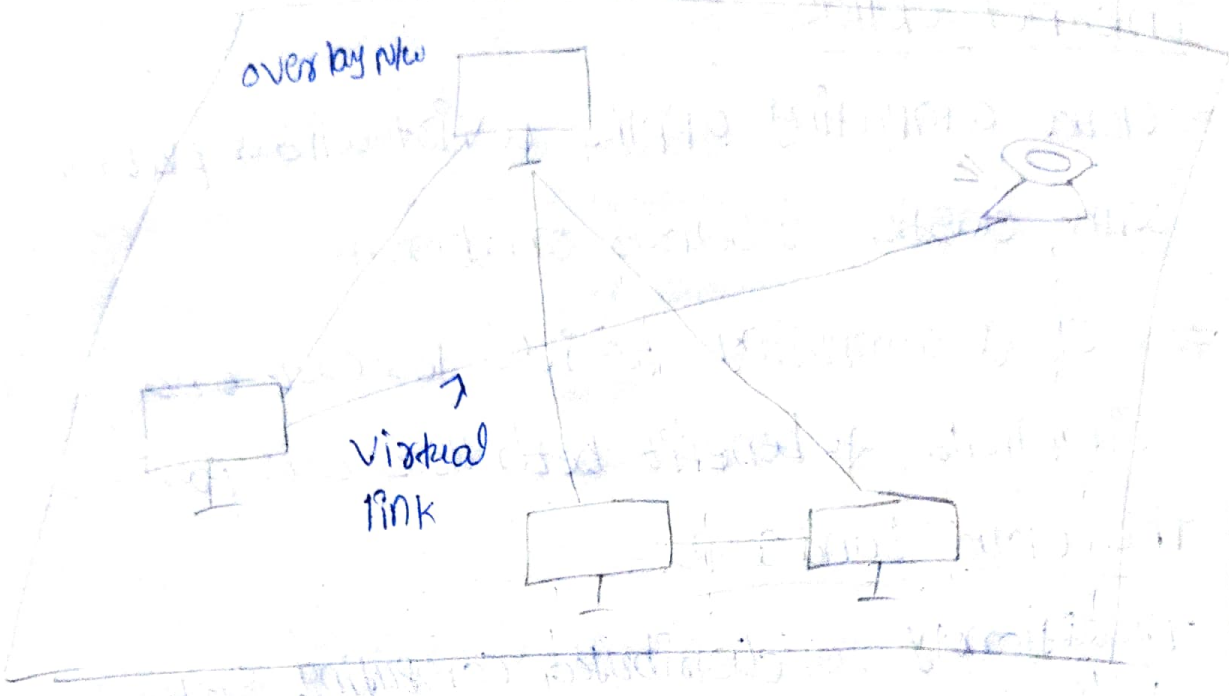
Peer-to-Peer network families:-

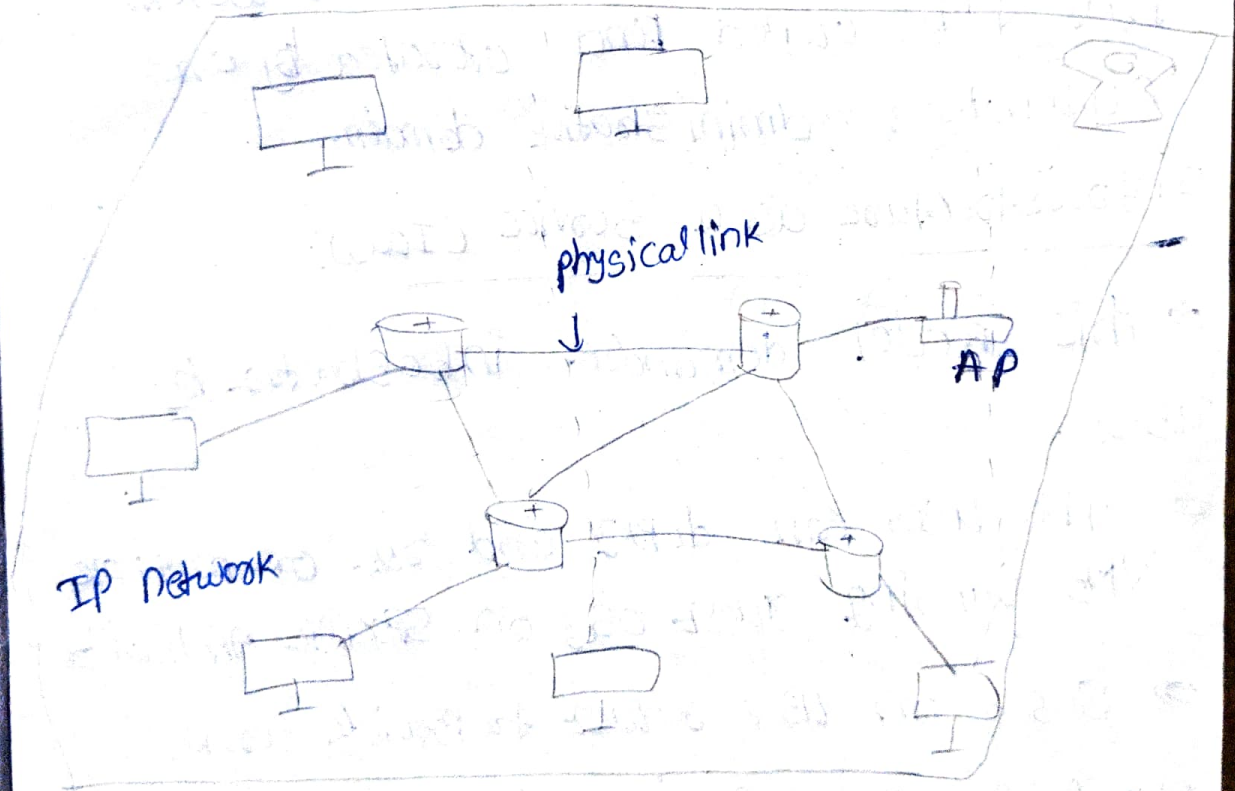
⇒ client machines are connected to a central server for compute, e-mail, file Access and database Applications

⇒ P2P network is client-oriented instead of server-oriented

overlay Networks:-

⇒ Data items are distributed in the participating peers





Cloud computing over the internet:-

⇒ A cloud is a pool of virtualized computer resources. A cloud can host a variety of different workloads, including batch-style backend jobs and interactive and user-facing applications.

Internet clouds:-

⇒ Cloud computing applies a virtualized platform with elastic resources on demand.

⇒ Cloud computing provide low cost and simplicity to benefit both users and providers.

The Cloud Landscape:-

Traditionaly, a distributed computing system tends to be owned and operated by an autonomous administrative domain.

Infrastructure as a Service (IaaS).

⇒ This model demanded infrastructure by users.

⇒ The user can deploy and run on multiple VMs running guest OSes on specific Applications.

⇒ Based on user request to provide services

Platform as a service (PaaS).

⇒ This model enables the user to deploy user built Applications onto a virtualized Cloud Platform.

⇒ The platform includes both software and hardware

⇒ The provider ~~the~~ supplies the API and software tools.

Software as a Service (SaaS):

⇒ This refers to browser-initiated application software over thousand of paid cloud customers.

⇒ It is applied business process, industry application, consumer relationship management.

Performance, Security and Energy efficiency

Performance metrics:-

⇒ Performance of a distributed system is measured by considering various aspects of the system such as processor speed, network bandwidth, response time etc.

⇒ In some cases, it can be categorized into two classes

1. Metrics associated with throughput
2. metrics associated with system availability.

⇒ Throughput is nothing but the total number of tasks that can be carried out by the system within a specified amount of time.

⇒ The metric Associated with it is MIPS (Million instructions Per second).

⇒ there exists some other metrics of measurement system throughput which include Terra Floating-Point operation (TFlops) and Transactions Per second (TPS).

⇒ The measuring System throughput include Job response time and network latency.

⇒ Network latency can be defined as the time taken by a data packet to travel from source to destination.

Dimensions of Scalability

(i) Size scalability:-

Size scalability refers to the process of improving system performance by updating the system with respect to hardware.

(ii) Software Scalability:-

Software scalability refers to the upgrading of system with respect to software including operating system, various libraries, applications, programming environment etc.

(iii) Application Scalability:-

⇒ This type of scalability typically identifies size scalability of problem and machine with which increasing the size of machine can be replaced with increasing the problem space.

(iv) Technology Scalability:-

This type of scalability refers to the upgrading of system with respect to various technologies such as computer networking, storage technology etc

(a) Time/Generation:

It is important an in-depth study on the current and upgraded generation of the system.

(b) Space:

space refers to usage of energy and packaging of various components

(c) Heterogeneity:-

It is a important aspect various Hardware and software components integrated to a single system

Amdahl's Law:-

⇒ Consider a program the total execution time of 'T' minutes

⇒ The program has been partitioned for parallel execution

⇒ Assume that a fraction α of the code must be executed sequentially, called the sequential bottleneck.

⇒ Therefore, $(1-\alpha)$ of the code can be compiled for parallel execution by n processors

⇒ The total execution time of the program is calculated by $\alpha T + (1-\alpha)T/n$.

⇒ Where the first term is the sequential execution time and second term is the parallel execution time.

⇒ The I/O time is not inauded in Analysis

$$\text{Speed up} = S = T[\alpha T + (1-\alpha)T/n] = 1[\alpha + (1-\alpha)/n]$$

Fault Tolerance and System availability :-

⇒ HA (high availability) is desired in all clusters, grids, P2P networks, and cloud systems

⇒ A system is highly available if it has a long mean time to failure (MTTF) and short mean time to Repair (MTTR)

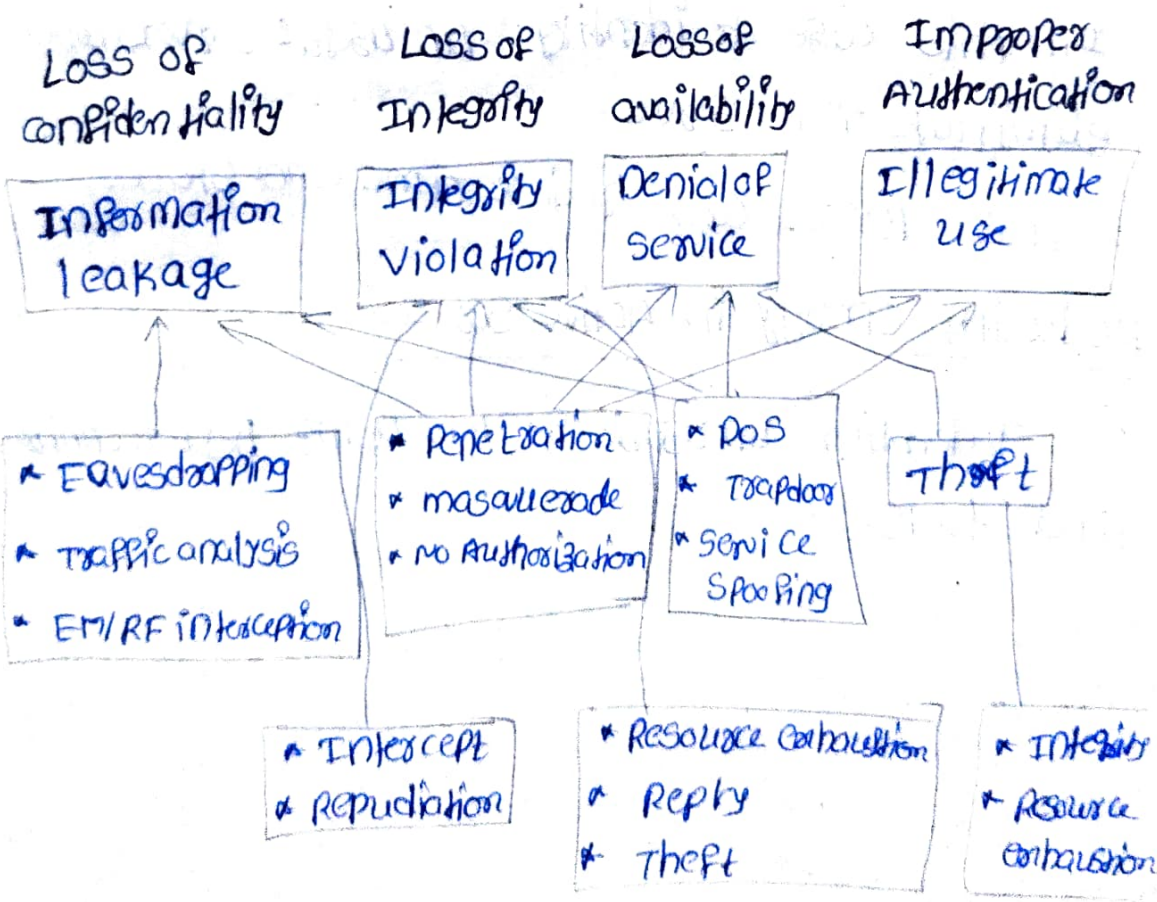
⇒ System availability is formally defined as

$$\text{system availability} = MTTF/(MTTF + MTTR)$$

⇒ Hardware, software and network components may fail.

⇒ Any failure that will pull down the operation of the entire system is called a single point of failure.

Network Threats and Data Integrity - Threats to systems and Networks:-



Copyright protection:-

⇒ Collusive piracy is the main source of intellectual property violations within the boundary of P2P n/w.

⇒ Pail clients (colluders) may illegally share copyrighted content files with unpaid clients (pirates).

Energy Efficiency in Distributed computing

⇒ primary performance goals in conventional parallel and distributed computing systems are high performence and high throughput.

Energy consumption of unused Servers:

⇒ Servers some are useful some are not useful.

In this case to identify un useful servers and eliminate this servers.

⇒ This process can reduce time and space.

Reducing energy in Active servers:

To Identify un usable system using best techniques and tools.