1 Network Computing

- In network computing, the nodes are stand-alone computers that could be connected
  - via a switch,
  - local area network,
  - the Internet.
- The main idea is to divide the application into semi-independent parts according to the kind of processing needed.
- Different nodes on the network can be assigned different parts of the application.
- Links are TCP/IP packet-switched connections and the bandwidth varies with load, number of hops, and underlying communication technology.
- Physical layers introduce delays and may be errors, which must be corrected by retransmission and dynamic reconfiguration of the Internet's links.

1.1 Computer Networks Basics

- The overall performance of a cluster system can be determined by
  1. the speed of its processors and
  2. the interconnection network.
- Many researchers argue that the interconnection network is the most important factor that affects cluster performance.
- Regardless of how fast the processors are, communication among processors, and hence scalability of applications, will always be bounded by the network bandwidth and latency.
  - **Bandwidth** is an indication of how fast a data transfer may occur from a sender to a receiver.
  - **Latency** is the time needed to send a minimal size message from a sender to a receiver.

Table 1: Data Rate, Switching Method, and Routing Scheme for Interconnection Networks.

Interconnection	Switching	Routing
Ethernet	Packet	Table-based
Fast Ethernet	Packet	Table-based
Gigabit Ethernet	Packet	Table-based
Myrinet	Wormhole	Source-path
Quadrics	Wormhole	Source-path
Infiniband	Packet	Source-path

- Networks can be divided into the following four categories based on their sizes and the geographic distances they cover:

1 **Wide area network (WAN)**; a WAN connects a large number of computers that are spread over large geographic distances. It can span sites in multiple cities, countries, and continents.

2 **Metropolitan area network (MAN)**; the MAN is an intermediate level between the LAN and WAN and can perhaps span a single city.

3 **Local area network (LAN)**; a LAN connects a small number of computers in a small area within a building or campus.

4 **System or storage area network (SAN)**; a SAN connects computers or storage devices to make a single system.

- In the early days of clusters, Ethernet was the main interconnection network used to connect nodes.

- While Ethernet resides at the low end of the performance spectrum, it is considered a low-cost solution.

- Other solutions add communication processors on the network interface cards, which provide programmability and performance.

- Table 1 shows the relative performance and other features of different high-speed networks.

- The major factor that distinguishes WAN from other network types is the scalability factor.

- LAN technologies provide higher speed connections compared to WAN because they cover short distances and hence offer lower delay than WANs.

- Network routing schemes can be classified as

- **connection-oriented**; in connection-oriented, the entire message follows the same path from source to destination.

  - Only the first packet holds routing information such as the destination address.

- **connectionless**; in connectionless schemes, a message is divided into packets.

  - The packets of a given message may take different routes from source to destination.
  - Therefore, the header of every packet holds routing information.
  - Using a serial number, the message can be reassembled in the correct order at the destination as packets may arrive in a different order.

1.1.1 Network Performance

- The following are two popular laws that predict the advances in network technologies.

1 **Gilder's Law**; George Gilder projected that the total bandwidth of communication systems triples every 12 months.

- Tells us that networking speed is increasing faster than processing power.

- While this remains true for the backbone network, end-to-end performance is likely to be limited by bottlenecks.

- For example, over about 15 years, LAN technology has increased in speed from 10 Megabits per second (10 Mbps) to 10 Giga-bits per second (10 Gbps), which is a factor of 1000 increase.

- Over a similar time period, advances in silicon technology, driven by Moore's Law, have allowed the CPU clock frequency in an average PC to increase from roughly 25 MHz to 2.5 GHz (a *factor of about 100 increase* in processing power).

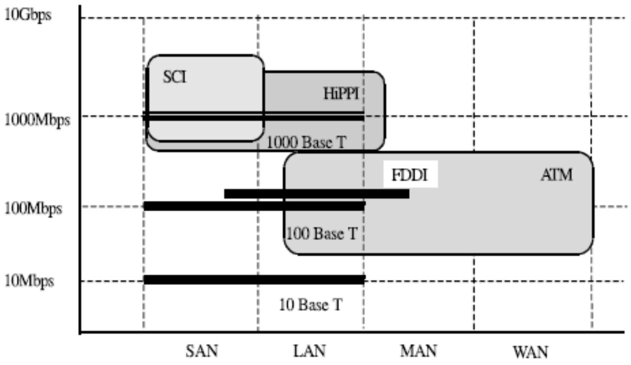

Figure 1: Representation of network technologies.

- *Metcalfe's Law*; Robert Metcalfe projected that the value of a network is proportional to the square of the number of nodes.

- Metcalfe's law also explains the productive growth of the Internet.

- As a network grows, the value of being connected to it grows exponentially, while the cost per user remains the same or even reduces.

- Internet is the collection of networks and routers that form a single co-operative virtual network, which spans the entire globe.

- The Internet relies on the combination of the Transmission Control Protocol and the Internet Protocol or TCP/IP.

- The majority of Internet traffic is carried using TCP/IP packets.

- With the projections of Gilder and Metcalfe, the number of users is expected to grow even more.

1.1.2 Other Network Technologies

- In addition to the popular TCP/IP protocol, many more protocols and combinations of protocols exist.

- Figure 1 shows different network technologies and their speed in relation to the network taxonomy.

- *Fast Ethernet and Gigabit Ethernet;*

- *The Fiber Distributed Data Interface (FDDI);*

  - The FDDI specifies a 100 Mbps token-passing, dual-ring LAN using fiber-optic cable.
  - The FDDI is frequently used as high-speed backbone technology because of its support for high bandwidth and greater distances than copper.

- *High-Performance Parallel Interface (HiPPI);*
  - The HiPPI is a point-to-point communication channel and it does not support multidrop configurations.
  - HiPPI is capable of transferring data at 800 Mbps using 32 parallel line or 1.6 Gbps over 64 parallel lines.

- *Asynchronous Transfer Mode (ATM);*
  - The ATM is a connection-oriented scheme that is suitable for both LANs and WANs.
  - It transfers data in small fixed-size packets called cells.
  - It can handle multimedia in an integrated way.
  - Cells are allowed to transfer using several different media such as both copper and fiberoptic cables.
  - It is designed to permit high-speed data. The fastest ATM hardware can switch data at a gigabit rate.

- *Scalable Coherent Interface (SCI);*
  - The SCI is an IEEE standard that is quite popular for PC clusters.
  - It represents a point-to-point architecture with directory-based cache coherence.
  - It provides a cluster-wide shared memory system.
  - A remote communication in SCI takes place as just part of a simple load or store process in a processor.

1.2 Client/Server Systems

- A Client/Server is a distributed system whereby the application is divided into at least two parts:

1 one or more servers perform one part

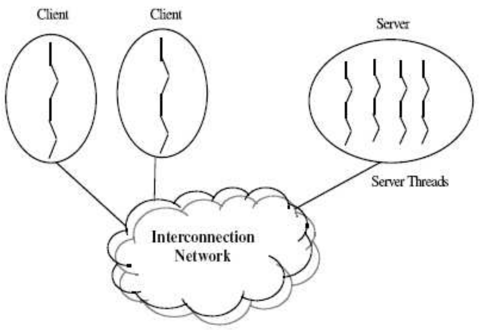

Figure 2: A multithreaded server in a client server system.

2 the other part is performed by one or more clients.

- Modern programming languages provide constructs for building client/server-based distributed applications.
- These applications are divided into clients and servers, which are allocated to different computers in a network.
  - A client sends a request to the server and waits for a response.
  - At the other end, when the server receives a request, it processes it and sends the results back to the client.
- In a database system, several clients send queries to the server that has access to the database.
- The server executes the queries on behalf of the clients and sends each client its respective result.
- A multithreaded process is considered an efficient way to provide server applications.
- A server process can service a number of clients as shown in Fig. 2.
- Each client request triggers the creation of a new thread in the server.

1.2.1 Sockets

- Sockets are used to provide the capability of *making connections from one application running on one machine to another running on a different machine.*

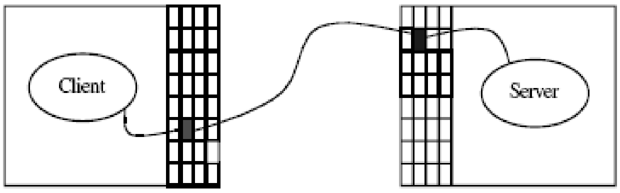

Figure 3: A socket connection.

- A socket abstraction consists of
  - the data structure that holds the information needed for communication,
  - the system calls that manipulate the socket structure.
- Once a socket is created, it
  - can be used to wait for an incoming connection (passive socket),
  - can be used to initiate connection (active socket).
- A client can establish an active connection to a remote server by creating an instance of a socket.
- A server socket listens on a TCP port for a connection from a client (passive socket).
- When a client connects to that port, the server accepts the connection (see Fig. 3).
- Once the connection is established, the client and server can read from and write to the socket using input and output streams.

1.2.2 A Client Server Framework for Parallel Applications

- Parallel applications can be designed using the client/server model.
- A client may divide a big application into several smaller problems that can be processed by multiple servers simultaneously.
- All the servers compute the solution to their respective problems and send their results to the client.

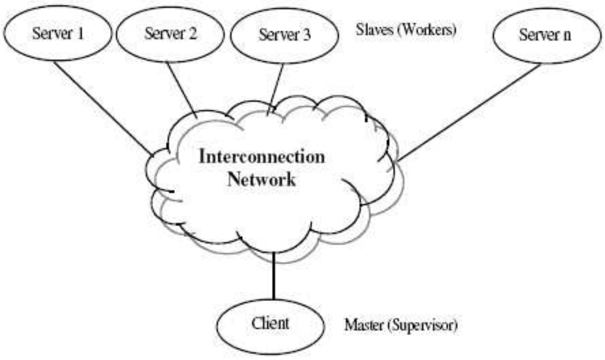

Figure 4: Supervisor workers model in client server.

- The client assembles the results from each server and outputs the final result to the user.
- The client acts as the master (supervisor) while the servers act as the slaves (workers) in the master-slave (supervisor-workers) model as shown in Fig. 4.

1.3 Clusters

- The 1990s have witnessed a significant shift from *expensive and specialized* parallel machines to the more cost-effective clusters of PCs and workstations.
- Advances in network technology and the availability of low-cost and high-performance commodity workstations have driven this shift.
- Clusters provide an economical way of achieving high performance.
- **Each node in a cluster could be a workstation, personal computer, or even a multiprocessor system**.
- **Each node has its own input/output systems and its own operating system.**
- When all nodes in a cluster have the same architecture and run the same operating system,
- the cluster is called **homogeneous**, otherwise, it is **heterogeneous**.

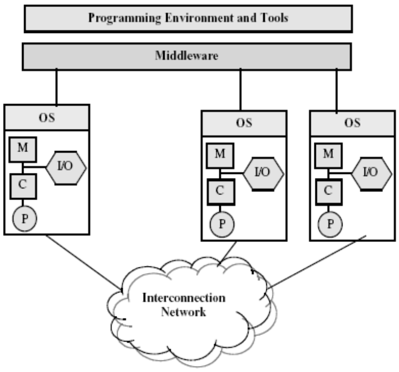

Figure 5: A cluster made of homogenous single-processor computers.

- Dedicated clusters are normally packaged compactly in a single room.
- With the exception of the front-end node, all nodes are headless with no keyboard, mouse, or monitor.

  The programming environment and tools layer provide the programmer with portable tools and libraries for the development of parallel applications.
- To achieve high-performance computing, the interconnection network must provide high-bandwidth and low-latency communication.
- Alternatively, nodes owned by different individuals on the Internet could participate in a cluster only part of the time.
- In this case, the cluster can utilize the idle CPU cycles of each participating node if the owner's permission is granted.
- The middleware layer in the architecture makes the cluster appears to the user as a single parallel machine, which is referred to as the single system image (SSI).
- The SSI infrastructure offers unified access to system resources by supporting a number of features including:
  - **Single entry point**: A user can connect to the cluster instead of to a particular node.
  - **Single file system**: A user sees a single hierarchy of directories and files.

  - **Single image for administration**: The whole cluster is administered from a single window.
  - **Coordinated resource management**: A job can transparently compete for the resources in the entire cluster.

- In addition to providing **high-performance computing**, clusters can also be used to provide **high-availability** environment.

- High availability can be achieved when only a subset of the nodes is used in the computation and the rest is used as a *backup in case of failure.*

- In cases when one of the main objectives of the cluster is high availability, the middleware will also support features that enable the cluster services for recovery from failure and fault tolerance among all nodes of the cluster.

- For example, the middleware should offer the necessary infrastructure for checkpointing.

- A checkpointing scheme makes sure that the process state is saved periodically.

- In the case of node failure, processes on the failed node can be restarted on another working node.

1.3.1 Cluster Examples

- The Berkeley Network of Workstations (NOW) is an important representative of cluster systems.

- In 1997, the NOW project achieved over 10 Gflops on the Linpack benchmark, which made it one of the top 200 fastest supercomputers in the world.

- The hardware/software infrastructure for the project included 100 SUN Ultrasparcs and 40 SUN Sparcstations running Solaris, 35 Intel PCs running Windows NT or a PC Unix variant, and between 500 and 1000 disks, all connected by a Myrinet switched network.

- The programming environments used in NOW are sockets, MPI, and a parallel version of C, called Split C.

- Active Messages is the basic communication primitive in Berkeley NOW.

- The idea of the Beowulf cluster project was to achieve supercomputer processing power using **off-the-shelf commodity** machines.

- One of the earliest Beowulf clusters contained sixteen 100 MHz DX4 processors that were connected using 10 Mbps Ethernet.

- The second Beowulf cluster, built in 1995, used 100 MHz Pentium processors connected by 100 Mbps Ethernet.

- The third generation of Beowulf clusters was built by different research laboratories JPL and Los Alamos National Laboratory each built a 16-processor machine incorporating Pentium Pro processors.

- These machines were combined to run a large N-body problem, which won the 1997 Gordon Bell Prize for high performance.

- The communication between processors in Beowulf has been done through TCP/IP over the Ethernet internal to the cluster.

- Multiple Ethernets were also used to satisfy higher bandwidth requirements.

- Channel bonding is a technique to connect multiple Ethernets in order to distribute the communication traffic.

- Channel bonding was able to increase the sustained network throughput by 75% when dual networks were used.

- Two of the early successful Beowulf clusters are Loki and Avalon.

- In 1997, Loki was built using 16 Pentium Pro Processors connected using Fast Ethernet switches. It achieved 1.2 Gflops.

- In 1998, the Avalon was built using one hundred and forty 533 MHz Alpha Microprocessors connected. Avalon achieved 47.7 Gflops.

- In April 2004, the University of San Francisco hosted the first Flash Mob Computing computer; FlashMob I, with the purpose of creating one of the fastest supercomputers on the planet.

- A FlashMob supercomputer was created by connecting a large number of computers via a high-speed LAN, to work together as a single supercomputer.

- A FlashMob computer, unlike an ordinary cluster, is temporary and organized *ad hoc* for the purpose of working on a single problem.
- It used volunteers and ordinary laptop PCs, and was designed to allow anyone to create a supercomputer in a matter of hours.
- Over 700 computers came into the gym and they were able to hook up 669 to the network.
- The best Linpack result was a peak rate of 180 Gflops using 256 computers; however, a node failed 75% through the computation.
- The best completed result was 77 Gflops using 150 computers.