

Multicomputer

- Multiple (full) computers connected by network.
- Distributed memory each have special address space.
- Access to data another processor is explicit in program, express by call function for sending or receiving message.
- Don't need operating System, enough libraries with function for sub sending message.
- Good scalability.

In this section we discuss network computing, in which the nodes are stand-alone computers that could be connected via a switch, local area network, or 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. This form of network computing takes advantage of the unique capabilities of diverse system architectures.

It also maximally leverages potentially idle resources within a large organization. Therefore, unused CPU cycles may be utilized during short periods of time resulting in bursts of activity followed by periods of inactivity. In what follows, we discuss the utilization of network technology in order to create a computing infrastructure using commodity computers.

Cluster

- In 1990 shifted from expensive and specialized parallel machines to the more cost-effective clusters of PCs and workstations.
- A cluster is a collection of stand-alone computers connected using some interconnection network.
- Each node in a cluster could be a workstation.
- Important for it to have fast processors and fast network to enable it to use for distributed system.
- Cluster workstation component:
- Fast processor/memory and complete HW for PC.
- Free access SW.
- High execute, low latency.

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. Departments that could not afford the expensive proprietary supercomputers have found an affordable alternative in clusters.

A cluster is a collection of stand-alone computers connected using some interconnection network. Each node in a cluster could be a workstation, personal computer, or even a multiprocessor system. A node is an autonomous computer that may be engaged in its own private activities while at the same time cooperating with other units in the context of some computational task. 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. The interconnection network could be a fast LAN or a switch. To achieve high-performance computing, the interconnection network must provide high-bandwidth and low-latency communication.

The nodes of a cluster may be dedicated to the cluster all the time; hence computation can be performed on the entire cluster. 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. Dedicated clusters usually use high-speed networks such as fast Ethernet and Myrinet such as (beowulf). 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.

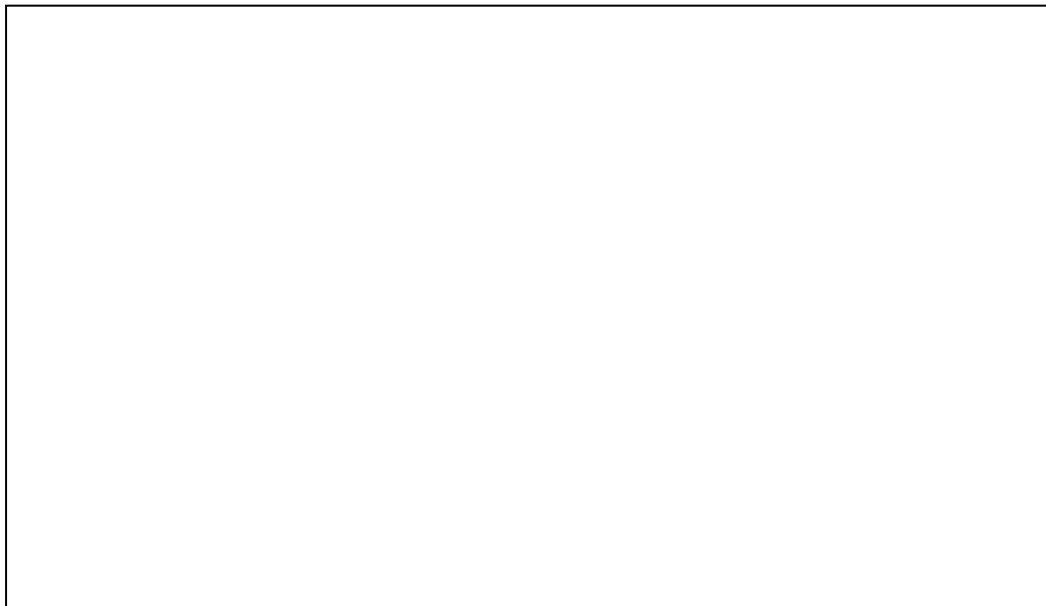
Beowulf

- The idea of the Beowulf cluster project was to achieve supercomputer processing power using off-the-shelf commodity machines.
- Many generations in ~1991,1995,1997....etc.
- Nodes are received only for distributed computing.
- Computer and network don't have another use (isolate cluster).

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

The nodes of a Beowulf dedicated to the cluster all the time; hence computation can be performed on the entire cluster. 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. Dedicated clusters usually use high-speed networks such as fast Ethernet and Myrinet.

Communication in data center



- **LAN**
 1. Connect various workstation and users.
 2. Connect to another network.
- **IPC(Inter-process Connect)**
 1. Communicate among nodes – processors.
 2. Serve to share message between executed nodes.
 3. Small volume data –low latency.
- **SAN (Storage Access Network)**
 1. Connect data storage.
 2. Disk feed share file.
 3. Big volume-big access.

Latency in computer Network

- Latency: time transfer empty message.
- Delay resources:
- Working over highest layers of (TCP/IP).
- Share system for send over media.
- Delay from transfer signal in media.
- Delay on switch:
 - Store and forward: LIFO, Cut-through: FIFO.

Minimize Latency

- Working with TCP/IP protocol on HW level:
- TCP/IP Slow initialize.
- TOE (TCP/IP Offload Engine): 3handshaking,ACK,seq numbers, Check sum, RST, Sliding windows).
- Use fast media:
- Some sharing media for short rang.
- Optical.
- Nonblocking “cut-through” switch.

Special technology

- **InfiniBand (IBTA):**
- Transfer rate one link: 2,4, 8 Gb/s, links is can connect to 1× ,4× or 12× (max. 96 Gb/s).
- Latency: 140ns, according specification.
- Big blocks.
- Establish “point-to-point” connection.
- Transfer media: copper wire, Optic.
- Suitable for IPC or SAN.

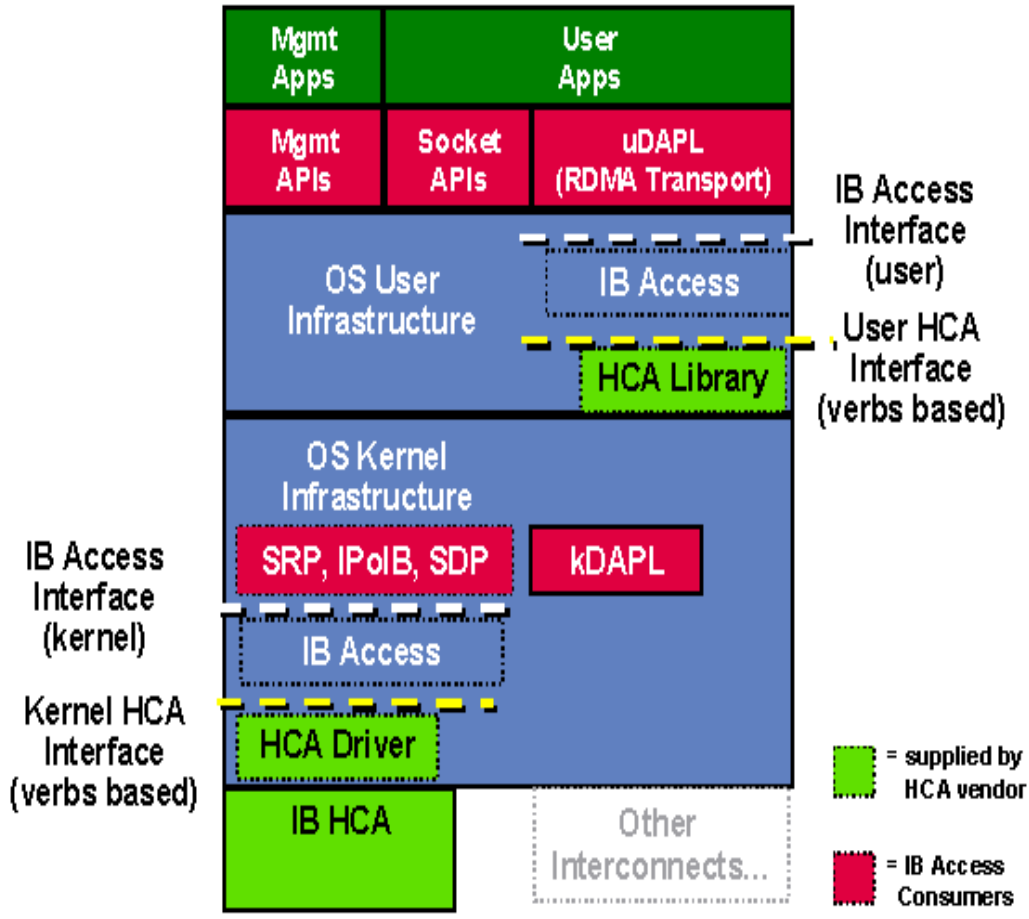
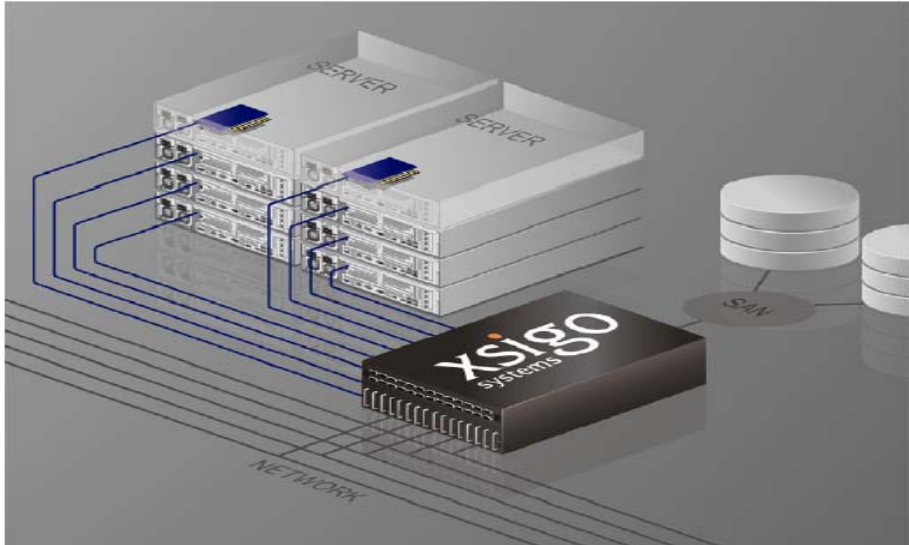
The InfiniBand® Trade Association was founded in 1999 and is chartered with maintaining and furthering the InfiniBand specification. The IBTA is led by a distinguished steering committee that includes IBM, Intel, Mellanox, QLogic, Sun and Voltaire. Other members of the IBTA represent leading enterprise IT vendors who are actively contributing to the advancement of the InfiniBand specification.

The InfiniBand™ Architecture (IBA) is an industry standard that defines a new high-speed switched fabric subsystem designed to connect processor nodes and I/O nodes to form a system area network. This new interconnect method moves away from the local transaction-based I/O model across busses to a remote message-passing model across channels. The architecture is independent of the host operating system (OS) and the processor platform.

IBA provides both reliable and unreliable transport mechanisms in which messages are enqueued for delivery between end systems. Hardware transport protocols are defined that support reliable and unreliable messaging (send/receive), and memory manipulation semantics (e.g., RDMA read/write) without software intervention in the data transfer path.

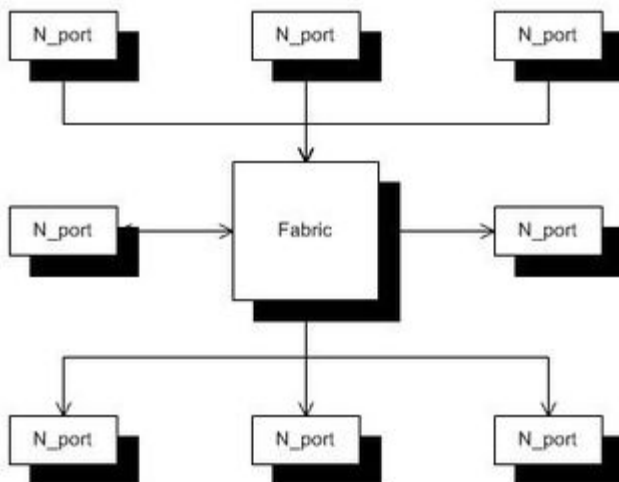
The InfiniBand specification primarily defines the hardware electrical, mechanical, link-level, and management aspects of an InfiniBand fabric, but does not define the lowest layers of the operating system stack needed to communicate over an InfiniBand fabric. The remainder of the operating system stack to support storage, networking, IPC, and systems management is left to the operating system vendor for definition.

<http://infiniband.sourceforge.net/>



- **Fiber channel (standard):**
- Transmission rate 2 GB/s till 4 GB/s.
- Transfer media: Optical.
- Establish “point-to-point” connection.
- Use for data storage (big volume data) (SAN).

Fibre Channel, or **FC**, is a gigabit-speed network technology primarily used for storage networking. Fibre Channel is standardized in the T11 Technical Committee of the InterNational Committee for Information Technology Standards (INCITS), an American National Standards Institute (ANSI)–accredited standards committee. It started use primarily in the supercomputer field, but has become the standard connection type for storage area networks (SAN) in enterprise storage. Despite its name, Fibre Channel signaling can run on both twisted pair copper wire and fiber-optic cables.

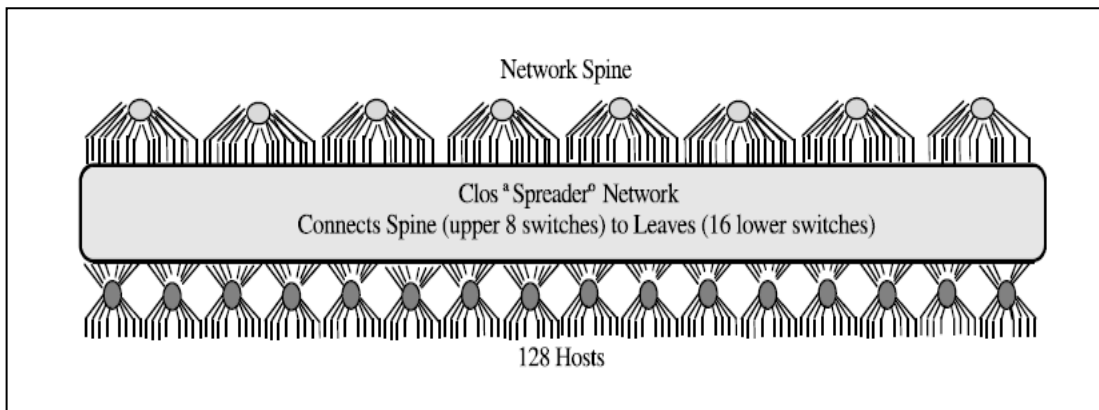


http://en.wikipedia.org/wiki/Fibre_Channel

- **Myrinet (from Myricom):**
- Transmission rate 10Gb/s (4th generation).
- Transfer media :2×Optical (upstream/ downstream).
- Low expensive.
- Use for LAN.

Myrinet is a high-performance, packet-communication and switching technology. It was produced by Myricom as a high-performance alternative to conventional Ethernet networks. Myrinet switches are multiple-port components that route a packet entering on an input channel of a port to the output channel of the port selected by the packet. Myrinet switches have 4, 8, 12, 16 ports. For an n-port switch, the ports are addressed 0, 1, 2, . . . , n - 1. For any switching permutation, there may be as many packets traversing a switch concurrently as the switch has ports.

The routing of Myrinet packets is based on the source routing approach. Each Myrinet packet has a variable length header with complete routing information. When a packet enters a switch, the leading byte of the header determines the outgoing port before being stripped off the packet header. At the host interface, a control program is executed to perform source-route translation.



Distribution share memory

Distributed Shared Memory (DSM), also known as a distributed global address space (DGAS), is a term in computer science that refers to a wide class of software and hardware implementations, in which each node of a cluster has access to a large shared memory in addition to each node's limited non-shared private memory.

- Combination sending message (in low level) and sharing memory in one address space.
- Benefit for application use one address space, which generate by sending message on low level (OS,HW).
- Memory is physically distributed, high level generate shared address space.
- Sending message is transparent.
- Require fast network.
- Virtual sharing memory:
- Local memory: table pages.
- Separate away memory: sending message.
- Coherence to safeguard all pages, big granularity, group communication.
- Objective distributed sharing memory:
- Data with access function (generate, access and modification).
- Is not transparent for programmer.

Hybrid system

A hybrid system structure comprised of distributed systems to take advantage of locality of reference and a central system to handle transactions that access non-local data is examined. Several transaction processing applications, such as reservation systems, insurance and banking have such regional locality of reference. A concurrency and coherency control protocol that maintains the integrity of the data and performs well for transactions that access local or non-local data is described. It is shown that the performance of the hybrid system is much less sensitive to the fraction of remote accesses than the distributed system and offers similar performance to the distributed system for local transactions.

- Using both, access to parts shared memory and sending message.
- Typically go about clusters of multiprocessors (multicomputer of multiprocessors).
- Ex. 16 computer in network, each has on his board 4 processors.
 - 4 processors for each node have a share memory.
 - Data between nodes is transfer by sending message.

Classification architecture

- According Flynn : SISD, MISD, SIMD, MIMD.
- According connect memory-processors:
 - 1- Multiprocessors (Share memory: UMA, NUMA, COMA)
 - 2- Multicomputer (dis. Memory: Sending message)
Cluster, Beowulf, grid.
 - 3- Distributed sharing memory(virtual share).
 - 4-hybrid system (content all type).
- According connection system:
 - 1- Static / dynamic.
 - 2- Direct / indirect sending.
 - 3- Type topology: linear, circular, star