Interconnection Networks and Clusters

Computer Architecture: A Quantitative Approach, 3rd Edition

Patterson and Hennessy
(Morgan Kaufmann Publishers, 2003)

Iqbal Syamsu (23205037)

Interconnection Networks and Clusters

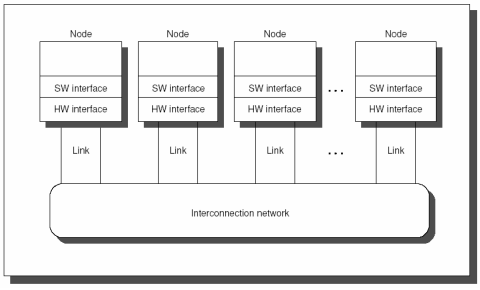
Introduction

Two reasons that computer architects should devote attention to networking:

- Computer architects must understand networking terminology, problems and solutions in order to design and evaluate modern computers.
- Today almost all computers are--or will be--networked to other devices. Thus, understanding networking is critical.

Interconnection Networks and Clusters

Introduction



Drawing of the generic interconnection network.

Interconnection Networks and Clusters

Introduction

Generic types of interconnections, depending on the number of nodes and their proximity:

- **Wide Area Network (WAN) / long haul network**

  connects computers distributed throughout the world. WANs include thousands of computers, and the maximum distance is thousands of kilometers. ATM is a current example of a WAN.

- **Local Area Network (LAN)**

  connects hundreds of computers, and the distance is up to a few kilometers. Unlike a WAN, a LAN connects computers distributed throughout a building or on a campus. The most popular and enduring LAN is Ethernet.

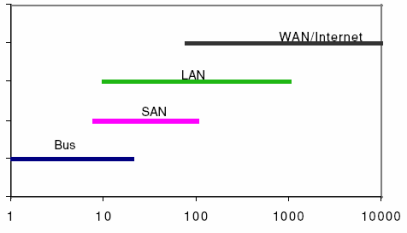
- **Storage or System area Network (SAN)**

  This interconnection network is for a machine room, so the maximum distance of a link is typically less than 100 meters, and it can connect hundreds of nodes. Today SAN usually means Storage area network as it connects computers to storage devices, such as disk arrays. Originally SAN meant a System area network to connect computers together, such as PCs in a cluster. A recent SAN trying to network both storage and system is Infiniband.

Interconnection Networks and Clusters

Introduction

Relationship of these systems in terms of number autonomous systems connected, including a bus for comparison.

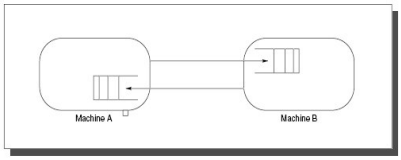


Relationship of four types of interconnects in terms of number of autonomous systems connected: bus, system or storage area network, local area network, and wide area network/Internet. Note that there are overlapping ranges where buses, SANs, and LANs compete. Some supercomputers have a switch-based custom network to interconnect up to thousands of computers; such interconnects are basically custom SANs.

Interconnection Networks and Clusters

Simple Network

Simple network connecting two machines.

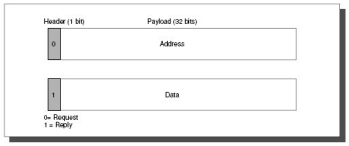


For one machine to get data from the other, it must first send a request containing the address of the data it desires from the other node. When a request arrives, the machine must send a reply with the data. Hence, each message must have at least 1 bit in addition to the data to determine whether the message is a new request or a reply to an earlier request. The network must distinguish between information needed to deliver the message, typically called the header or the trailer depending on where it is relative to the data, and the payload, which contains the data.

Interconnection Networks and Clusters

Simple Network

Message format for simple network, messages must have extra information beyond the data



The software steps to send a message are as follows:

1. The application copies data to be sent into an operating system buffer.
2. The operating system calculates the checksum, includes it in the header or trailer of the message, and then starts the timer.
3. The operating system sends the data to the network interface hardware and tells the hardware to send the message.

Interconnection Networks and Clusters

Simple Network

Message format for simple network, messages must have extra information beyond the data

Message reception is in just the reverse order:

1. The system copies the data from the network interface hardware into the operating system buffer.
2. The system calculates the checksum over the data. If the checksum matches the sender's checksum, the receiver sends an acknowledgment back to the sender. If not, it deletes the message, assuming that the sender will resend the message when the associated timer expires.
3. If the data pass the test, the system copies the data to the user's address space and signals the application to continue.

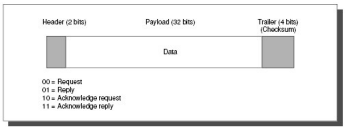
The sender must still react to the acknowledgment:

- When the sender gets the acknowledgment, it releases the copy of the message from the system buffer.
- If the sender gets the time-out instead of an acknowledgment, it resends the data and restarts the timer.

Interconnection Networks and Clusters

Simple Network

Message format for simple network, messages must have extra information beyond the data



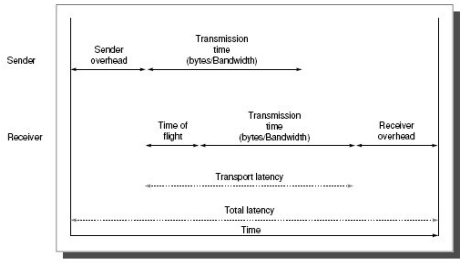
The sequence of steps that software follows to communicate is called a **protocol** and generally has the symmetric but reversed steps between sending and receiving.

Note that this example protocol above is for sending a single message. When an application does not require a response before sending the next message, the sender can overlap the time to send with the transmission delays and the time to receive.

Interconnection Networks and Clusters

Simple Network

Performance parameters of interconnection networks.



Depending on whether it is an SAN, LAN, or WAN, the relative lengths of the time of flight and transmission may be quite different from those shown here.
(Based on a presentation by Greg Papadopolous of Sun Microsystems.)

Interconnection Networks and Clusters

Simple Network

Confusing Terms

- **Bandwidth**

  We use this most widely used term to refer to the maximum rate at which the network can propagate information once the message enters the network. Unlike disks, bandwidth includes the headers and trailers as well as the payload, and the units are traditionally bits/second rather than bytes/second. The term bandwidth is also used to mean the measured speed of the medium or network bandwidth delivered to an application. Throughput is sometimes used for this latter term.

- **Time of Flight**

  The time for the first bit of the message to arrive at the receiver, including the delays due to repeaters or other hardware in the network. Time of flight can be milliseconds for a WAN or nanoseconds for an SAN.

Interconnection Networks and Clusters

Simple Network

Confusing Terms…cont (1)

- **Transmission Time**

  The time for the message to pass through the network, not including time of flight. One way to measure it is the difference in time between when the first bit of the message arrives at the receiver and when the last bit of the message arrives at the receiver. Note that by definition transmission time is equal to the size of the message divided by the bandwidth. This measure assumes there are no other messages to contend for the network.

- **Transport Latency**

  The sum of time of flight and transmission time. Transport latency is the time that the message spends in the interconnection network. Stated alternatively, it is the time between when the first bit of the message is injected into the network and when the last bit of the message arrives at the receiver. It does not include the overhead of injecting the message into the network nor pulling it out when it arrives.

Interconnection Networks and Clusters

Simple Network

Confusing Terms…cont (2)

- **Sender Overhead**

  The time for the processor to inject the message into the network, including both hardware and software components. Note that the processor is busy for the entire time, hence the use of the term overhead. Once the processor is free, any subsequent delays are considered part of the transport latency. For pedagogic reasons, we assume overhead is not dependent on message size. (Typically, only very large messages have larger overhead.)

- **Receiver Overhead**

  The time for the processor to pull the message from the interconnection network, including both hardware and software components. In general, the receiver overhead is larger than the sender overhead: for example, the receiver may pay the cost of an interrupt.

Interconnection Networks and Clusters

Simple Network

The total latency of a message

$$\text{Total latency} = \text{Sender overhead} + \text{Time of flight} + \frac{\text{Message size}}{\text{Bandwidth}} + \text{Receiver overhead}$$

Interconnection Networks and Clusters

Simple Network

Example

- **Problem**

  Assume a network with a bandwidth of 1000 Mbits/second has a sending overhead of 80 microseconds and a receiving overhead of 100 microseconds. Assume two machines. One wants to send a 10000-byte message to the other (including the header), and the message format allows 10000 bytes in a single message. Let's compare SAN, LAN, and WAN by changing the distance between the machines. Calculate the total latency to send the message from one machine to another in a SAN assuming they are 10 meters apart. Next, perform the same calculation but assume the machines are now 500 meters apart, as in a LAN. Finally, assume they are 1000 kilometers apart, as in a WAN.

Interconnection Networks and Clusters

Simple Network

Example

- **Answer**

  The speed of light is 299,792.5 kilometers per second in a vacuum, and signals propagate at about 63% to 66% of the speed of light in a conductor. Since this is an estimate, in this chapter we'll round the speed of light to **300,000 km/sec**, and assume we can achieve two-thirds of that in a conductor. Hence, we can estimate time of flight. Let's plug the parameters for the short distance of a SAN into the formula above:

Total Lat. = Tx overhead + Time of flight + (Message size/Bandwidth)+ Rx overhead

$$= 80\text{us} + [0.01\text{km} / (2/3 \times 30\text{x}10^5\text{km/s})] + 10000\text{byte}/(1000\text{MB/s}) + 100\text{us}$$

$$= 80\text{us} + 0.05\text{us} + 80\text{us} + 100\text{us}$$

$$= 260\text{us}$$

Interconnection Networks and Clusters

Simple Network

Example

- **Total Lat.** $_{(10m)}$ = 260 µsecs
- **Total Lat.** $_{(500m)}$ = 262 µsecs
- **Total Lat.** $_{(1000km)}$ = 5260 µsecs

As mentioned above, when an application does not require a response before sending the next message, the sender can overlap the sending overhead with the transport latency and receiver overhead. Increased latency affects the structure of programs that try to hide this latency, requiring quite different solutions if the latency is 1, 100, or 10,000 microseconds!

Interconnection Networks and Clusters

Simple Network

Simplify the performance equation

Time of flight for SANs is so short relative to overhead that it can be ignored, yet in WANs, time of flight is so long that sender and receiver overheads can be ignored.

- **Total Lat.** ≈ Overhead + $\frac{\text{Message Size}}{\text{Bandwidth}}$

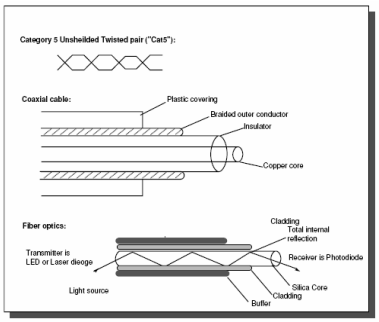
- **Effective bandwidth** = $\frac{\text{Message Size}}{\text{Total Latency}}$

Interconnection Networks and Clusters

Connecting More Than Two Computers

Three popular network media

- Twisted pairs of copper wires
- Coaxial cable
- Fiber optics



Optics:

- **Multimode fiber: inexpensive LEDs as a light source, 62.5 microns dia, 100~1000 Mb/s.**
- **Single-mode fiber: expensive, 8 to 9 microns dia, transmits gigabits per second for hundreds of kilometers.**

Interconnection Networks and Clusters

Connecting More Than Two Computers

Example

Problem

Suppose you have **25** magnetic tapes, each containing **40 GB**. Assume that you have enough tape readers to keep any network busy. How long will it take to transmit the data over a distance of one kilometer? Assume the choices are Category 5 twisted pair wires at **100 Mbits/second**, multimode fiber at **1000 Mbits/second**, and single mode fiber at **2500 Mbits/second**. How do they compare to delivering the tapes by car?

Answer : The amount of data is 1000 GB.

- **Twisted pair** = $\frac{1000 \times 1024 \times 8 \text{ Mb}}{100\text{Mb/sec}}$ = 81,920 secs = 22.8 hours
- **Multimode fiber** = $\frac{1000 \times 1024 \times 8 \text{ Mb}}{1000\text{Mb/sec}}$ = 8192 secs = 2.3 hours
- **Single-mode fiber** = $\frac{1000 \times 1024 \times 8 \text{ Mb}}{2500\text{Mb/sec}}$ = 3277 secs = 0.9 hours

Interconnection Networks and Clusters

Connecting More Than Two Computers

Example

Answer : *(cont)*

- **Car = Time to load car + Transport time + Time to unload car**

$$= 300 \text{ secs} + \frac{1 \text{ km}}{30\text{kph}} + 300 \text{ secs} = 300 \text{ secs} + 120 \text{ secs} + 300 \text{ secs}$$

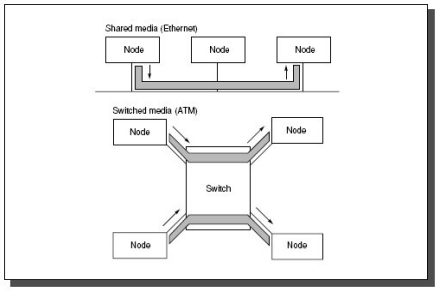
$$= 720 \text{ secs} = 0.3 \text{ hours}$$

A car filled with high-density tapes is a high-bandwidth medium!

Interconnection Networks and Clusters

Connecting More Than Two Computers

Shared medium versus switch



Ethernet was originally a shared medium, and but Ethernet switches are now available. All nodes on the shared media must share the 100 Mb/sec interconnection, but switches can support multiple 100 Mb/sec transfers simultaneously. Low cost Ethernet switches are sometimes implemented with an internal bus with higher bandwidth, but high-speed switches have a cross-bar interconnect.

Interconnection Networks and Clusters

Connecting More Than Two Computers

Example

Problem

Compare 16 nodes connected three ways: a single 100 Mb/sec shared media; a switch connected via Cat5, each segment running at 100 Mb/sec; and a switch connected via optical fibers, each running at 1000 Mb/sec. The shared media is 500 meters long, and the average length of each segment to a switch is 50 meters. Both switches can support the full bandwidth.

Assume each switch adds 5 microseconds to the latency. Calculate the aggregate bandwidth and transport latency. Assume the average message size is 125 bytes, and ignore the overhead of sending or receiving a message and contention for the network.

Answer

The aggregate bandwidth of each example is the simplest calculation: 100 Mb/sec for the shared media; 16 × 100, or 1600 Mb/sec for the switched twisted pairs; and 16 × 1000, or 16000 Mb/sec for the switched optical fibers.

Interconnection Networks and Clusters

Connecting More Than Two Computers

Example

Answer

$$\text{Transport time} = \text{Time of flight} + \frac{\text{Message size}}{\text{Bandwidth}}$$

Coax:

$$\text{Transport time}_{\text{SHARED}} = \frac{500/1000}{2/3 \times 300.000}\ 10^6\ \mu\text{secs} + \frac{125 \times 8}{100}\ \mu\text{secs}$$

$$= 2.5\ \mu\text{secs} + 10\ \mu\text{secs}$$

$$= 12.5\ \mu\text{secs}$$

Switch:

For the switches, the distance is twice the average segment, since there is one segment from the sender to the switch and one from the switch to the receiver. We must also add the latency for the switch.

Interconnection Networks and Clusters

Connecting More Than Two Computers

Example

- Answer

Switch:

$$\text{Transport time}_{SWITCH} = 2 \left(\frac{50/1000\ 10^6}{2/3 \times 300.000} \right) \mu s + 5\ \mu s + \frac{125 \times 8\ \mu s}{100}$$

$$= 0.5\ \mu secs + 5\ \mu secs + 10\ \mu secs$$

$$= 15.5\ \mu secs$$

Fiber:

$$\text{Transport time}_{FIBER} = 2 \left(\frac{50/1000\ 10^6}{2/3 \times 300.000} \right) \mu s + 5\ \mu s + \frac{125 \times 8\ \mu s}{1000}$$

$$= 0.5\ \mu secs + 5\ \mu secs + 1\ \mu secs$$

$$= 6.5\ \mu secs$$

Interconnection Networks and Clusters

Connecting More Than Two Computers

Connection-Oriented versus Connectionless Communication

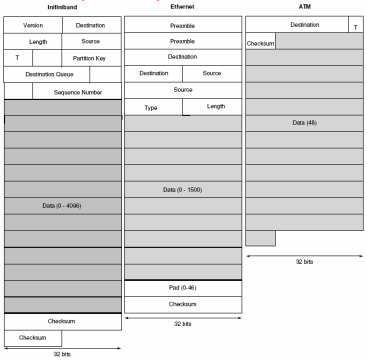
- Example

Transport

Interconnection Networks and Clusters

Examples of Interconnection Networks

Packet format for Infiniband, Ethernet, and ATM.

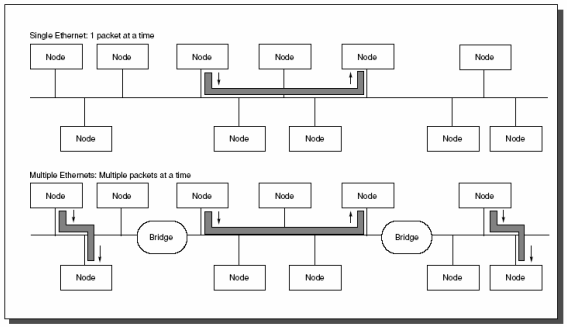


Interconnection Networks and Clusters

Examples of Interconnection Networks

Ethernet: The Local Area Network

- The potential increased bandwidth of using many Ethernets and bridges



Interconnection Networks and Clusters

Examples of Interconnection Networks

Devices

- Bridges

  Connect LANs together, passing traffic from one side to another depending on the addresses in the packet. Bridges operate at the Ethernet protocol level and are usually simpler and cheaper than routers

- Routers or gateways

  Connect LANs to WANs or WANs to WANs, and resolve incompatible addressing. Generally slower than bridges, they operate at OSI layer 3, the network layer. Routers divide the interconnect into separate smaller subnets, which simplifies manageability and improves security.

- Hubs

  Merely extend multiple segments into a single LAN. Thus, hubs do not help with performance, as only one message can transmit at a time. Hubs operate at OCI layer 1, the physical layer.

Interconnection Networks and Clusters

Examples of Interconnection Networks

Storage Area Network: Infiniband

- Infiniband

  A SAN that tries to optimize based on shorter distances

- Clock rates of 2.5 GHz
- Speed: 2000 Mbits/s per link
- Packet switched, connectionless network
- The distances are much shorter than Ethernet

  Cat-5 wire: 17m, optical fiber: 100 m

- Compared to LAN

  protocol overhead is much lower, protection is much more important in the LAN than the SAN, congestion is critical.

Interconnection Networks and Clusters

Examples of Interconnection Networks

Wide Area Network: ATM

- **Asynchronous Transfer Mode**

  155 Mbits/sec, and scales by factors of four to 620 Mbits/sec, 2480 Mbits/sec, and so on…

- **Small, fixed sized packet (makes it simpler to have fast routers and switches)**

Interconnection Networks and Clusters

Examples of Interconnection Networks

Summary

	LAN			SAN			WAN
	10-Mb Ethernet	100-Mb Ethernet	1000-Mb Ethernet	FC-AL	Infiniband	Myrinet	ATM
Length (meters)	500/2500	200	100	30/1000	17/100	10/550/10000	
Number data lines	1	1	4/1	2	1, 4, or 12	?	1
Clock rate (MHz)	10	100	1000	1000	2500	1000	155/622…
Switch?	Optional	Optional	Yes	Optional	Yes	Yes	Yes
Nodes	≤254	≤254	≤254	≤127	≤~1000	≤~1000	~10000
Media	Copper	Copper	Copper/fiber	Copper/fiber	Copper/fiber	Copper/m.m./s.m.fiber	Copper/fiber
Peak link BW (Mbits/sec)	10	100	1000	800	2000, 8000, or 24000	1300 to 2000	155/622/…
Topology	Line or Star	Line or Star	Star	Ring or Star	Star	Star	Star
Connec-tionless?	Yes	Yes	Yes	Yes	Yes	Yes	No
Routing	Dest. based	Dest. based	Dest. based	Destina-tion based	Destina-tion based	Dest. based	Virtual circuit
Store & forward?	No	No	No	No	No	No	Yes
Congestion control	Carrier sense	Carrier sense	Carrier sense	Credit-based	Back-pressure	Back-pressure	Credit based
Standard	IEEE 802.3	IEEE 802.3	IEEE 802.3ab-1999	ANSI Task Group X3T11	Infiniband Trade Association	ANSI/VITA 26-1998	ATM Forum

Interconnection Networks and Clusters

Examples of Interconnection Networks

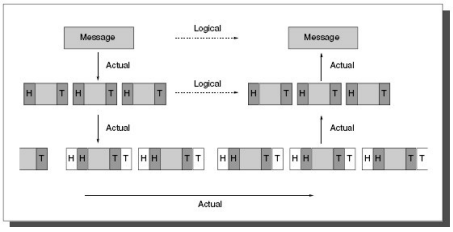
Summary (Open Systems Interconnect Layer)

Layer number	Layer name	Main Function	Example Protocol	Network component
7	Application	Used for applications specifically written to run over the network	FTP, DNS, NFS, http	Gateway, smart switch
6	Presentation	Translates from application to network format, and vice-versa		Gateway
5	Session	Establishes, maintains and ends sessions across the network	Named pipes, RPC	Gateway
4	Transport	Additional connection below the session layer	TCP	Gateway
3	Network	Translates logical network address and names to their physical address (e.g., computer name to MAC address)	IP	Router, ATM switch
2	Data Link	Turns packets into raw bits and at the receiving end turns bits into packets	Ethernet	Bridge, Network Interface Card
1	Physical	Transmits raw bit stream over physical cable	IEEE 802	Hub

Interconnection Networks and Clusters

Internetworking

A generic protocol stack with two layers

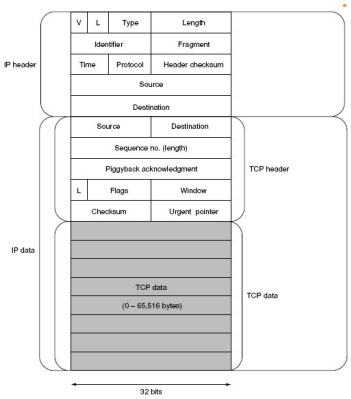


Note that communication is peer-to-peer, with headers and trailers for the peer added at each sending layer and removed by each receiving layer. Each layer offers services to the one above to shield it from unnecessary details

Interconnection Networks and Clusters

Internetworking

The headers for



Note that cc …d trailers for the peer ad… …ch receiving layer. Each … …d it from unnecessar…

Interconnection Networks and Clusters

Internetworking

Issues for Interconnection Networks

- **Density-Optimized Processors versus SPEC-optimized Processors**

  One microprocessor in 2001 burns 135 watts!

  SPEC -> Benchmark

- **Smart Switches vs. Smart Interface Cards**
- **Protection and User Access to the Network**
- **Compute-Optimized Processors versus Receiver Overhead**

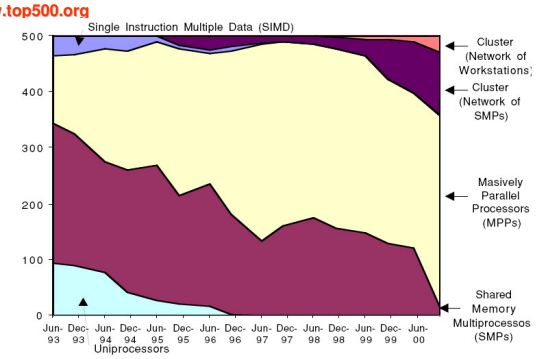
  missed instruction issue opportunities per message reception is likely to rise quickly over time.

Interconnection Networks and Clusters

Cluster

Plot of Top 500 supercomputer sites between 1993 and 2000

- www.top500.org

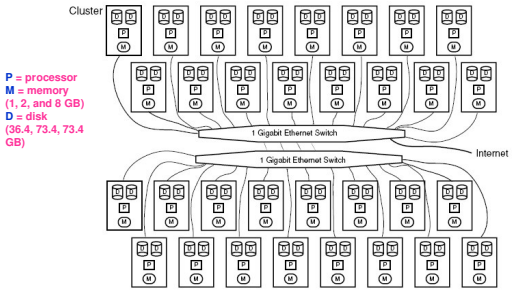


Interconnection Networks and Clusters

Cluster

Designing a Cluster

- Three cluster organizations.... based on uniprocessors

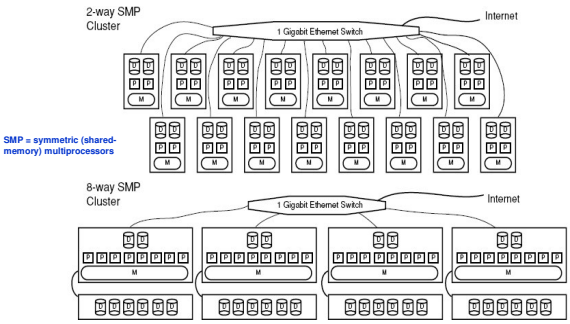


Interconnection Networks and Clusters

Cluster

Designing a Cluster

- Three cluster organizations.... 2-way SMPs, and 8-way SMPs

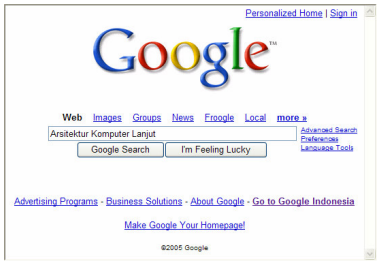


Interconnection Networks and Clusters

Cluster

Example

- Google Cluster



Interconnection Networks and Clusters

Cluster

Example

- Google Cluster

In December 2000 Google uses more than 6000 processors and 12000 disks, giving Google a total of about one petabyte of disk storage. At the time, the Google site was likely the single system with the largest storage capacity in the private sector.

2003: More than 15,000 commodity class pcs with fault-tolerant software.

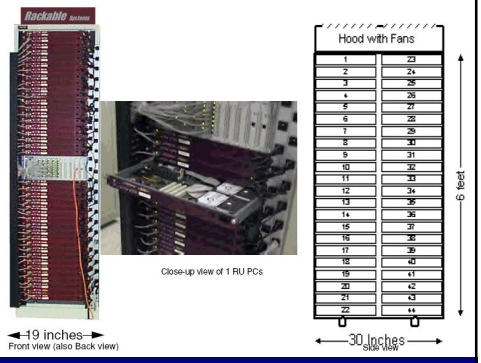
Interconnection Networks and Clusters

Cluster

Example

- Google Cluster

Front view, side view, and close-up of a rack of PCs used by Google.



Interconnection Networks and Clusters

Example

Google Cluster

The photograph on the left shows the HP Procurve 4000M Ethernet switch in the middle, with 20 PCs above and 20 PCs below. Each PC connects via a Cat5 cable on the left side to the switch in the middle, running 100 Mbit Ethernet. Each "blade" of the switch can hold 8 100 Mbit Ethernet interfaces or 1 1Gbit interface. There are also two 1 Gbit Ethernet links leaving the switch on the right. Thus, each PC has only 2 cables: 1 Ethernet and 1 power cord. The far right of the photo shows a power strip, with each of the 40 PCs and the switch connected to it. Each PC is 1 VME rack unit (RU) high. The switch in the middle is 4 RU high. The photo on the middle is a close up of rack, showing contents of a 1 RUPC. This unit contains 2 Maxtor DiamondMax 5400 RPM IDE drives on the right of the box, 256 MB of 100 MHz SDRAM, a PC motherboard, a single power supply, and an Intel microprocessor.

Each PC runs versions 2.2.16 or 2.2.17 Linix kernels on a slightly modified RedHat release. Between March 2000 and November 2000, over the period the Google site was populated, the microprocessor varied in performance from a 533 MHz Celeron to an 800 MHz Pentium III. The goal was selecting good cost performance, which was often close to $200 per chip. Disk capacity varied from 40 to 80 GB. You can see the Ethernet cables on the left, power cords on the right, and table Ethernet cables connected to the switch at the top of the figure. In December 2000 the unassembled parts costs are about $500 for the two drives, $200 for the microprocessor, $100 for the motherboard, and $100 for the DRAM. Including the enclosure, power supply, fans, cabling and so on, an assembled PC might cost $1300 to $1700. The drawing on the right shows that PCs are kept in two columns, front and back, so that a single rack holds 80 PCs and 2 switches. The typical power per PC is about 55 watts and about 70 watts per switch, so a rack uses about 4500 watts. Heat is exhausted into a 3-inch vent between the two columns, and the hot air is drawn out the top using fans. (The drawing shows uses 22 PCs per side each 2 RU high instead of the Google configuration of 40 1 RU PCs plus a switch per side.) (Photos and figure from Rackable Systems: http://www.rackable.com/advantage.htm).