High Bandwidth Network Switching

- The need for switching
  - the lessons of FDDI
  - planning for multimedia traffic
- ATM and switched Ethernet standards
  - how they work
  - cost

FDDI - History

FDDI is an ANSI standard and is available from:

American National Standards Institute

11 West 42nd Street

New York, New York 10036

Initial proposals for MAC and PHY submitted June 1983

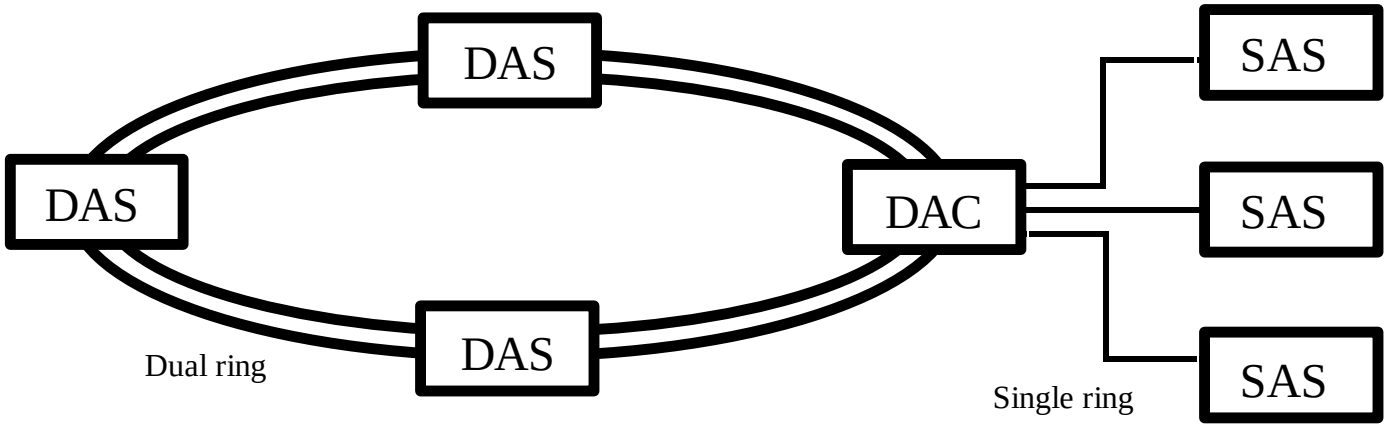
1. February 1986 MAC (Rev. 10) forwarded to X3
2. August 1985 PHY (Rev. 11) forwarded to X3
   Problem with specification of elasticity buffer discovered
3. August 1987 PHY (Rev. 15) forwarded to X3
4. June 1988 PMD (Rev. 8) forwarded to X3

ANSI X3.139-1987	FDDI Token Ring Media Access Control (MAC)
ANSI X3.148-1988	FDDI Token Ring Physical Layer Protocol (PHY)
ANSI X3.184-1993	FDDI Single-Mode Fiber Physical Layer Medium Dependent (SMF-PMD)
ANSI X3.231-1994	FDDI Token Ring Physical Layer Protocol (PHY-2)
ANSI X3.239-1994	FDDI Token Ring Media Access Control-2 (MAC-2)
ANSI X3.229-1994	FDDI Station Management (SMT)

FDDI - Introduction

FDDI is a token passing network with single or dual ring station interface implemented with fiber optics, shielded twisted pair, or unshielded twisted pair. The data rate is 100 Mb/s, default network parameters assume 1000 stations and 200 km circumference.
There are three basic types of station; single attach, dual attach, and dual attach concentrator.

A dual ring provides secondary backup in case of primary ring failure

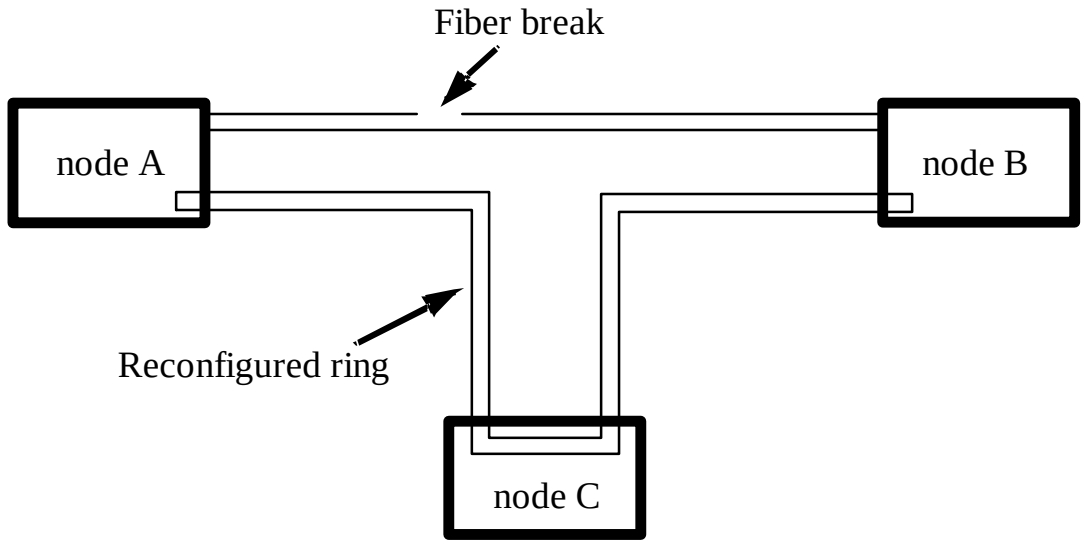


FDDI Topology

FDDI is a ring topology. Alternative solutions such as bus and star were either too difficult/expensive to implement at the physical layer or susceptible to failure.

Ring Topology: Consists of a series of point-to-point links with a bypass mechanism in case of single-node failure e.g. optical bypass

Dual Ring Topology: Adds redundancy to make network more reliable (no longer susceptible to fiber breaks)



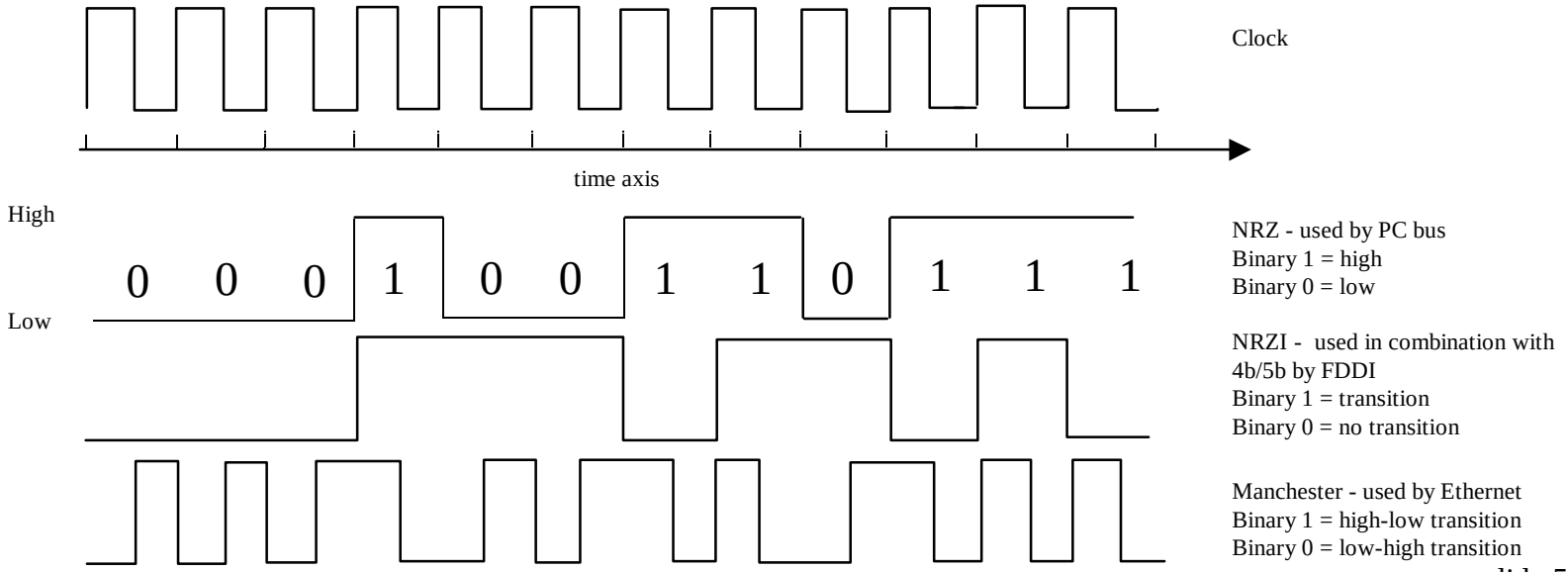
Signal Encoding

Signal encoding is used to increase system robustness against noise

Examples:

FDDI uses 4b/5b NRZI (Non-Return to Zero Invert on ones) with 125 Mb/s baud rate to achieve 100 Mb/s data rate

Ethernet uses Manchester encoding with 20 Mb/s baud rate (20 MBd) to achieve 10 Mb/s data rate

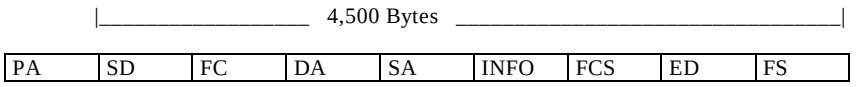


4b/5b Encoding

The 4b/5b code-bit stream has a maximum of three consecutive code-cell zeros with a maximum +/-10% cumulative dc component variation from nominal center.

This coding is convenient for ac coupling thereby reducing the noise component in receiver circuitry and is also useful for self-clocking.

FDDI uses 4b/5b coding combined with NRZI

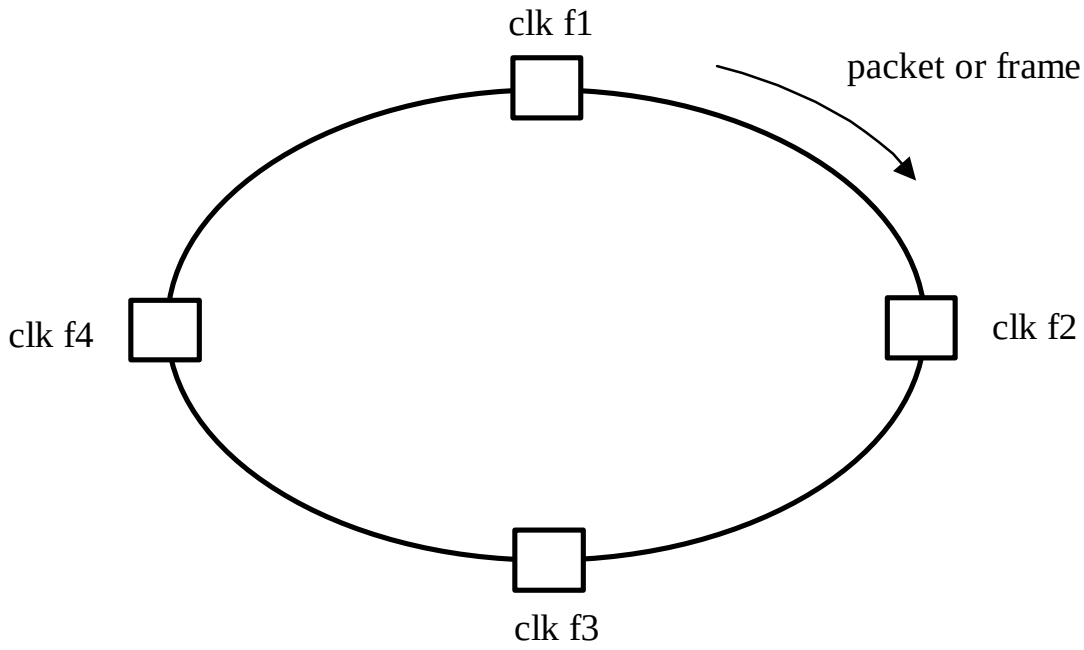


4.5 kB maximum packet size

Symbol	Bit Encoding	Meaning
0	11110	Data Bits, Value: 0x0, binary 0000
1	01001	Data Bits, Value: 0x1, binary 0001
2	10100	Date Bits, Value: 0x2, binary 0010
3	10101	Data Bits, Value: 0x3, binary 0011
4	01010	Data Bits, Value: 0x4, binary 0100
5	01011	Data Bits, Value: 0x5, binary 0101
6	01110	Data Bits, Value: 0x6, binary 0110
7	01111	Data Bits, Value: 0x7, binary 0111
8	10010	Data Bits, Value: 0x8, binary 1000
9	10011	Data Bits, Value: 0x9, binary 1001
A	10110	Data Bits, Value: 0xA, binary 1010
B	10111	Data Bits, Value: 0xB, binary 1011
C	11010	Data Bits, Value: 0xC, binary 1100
D	11011	Data Bits, Value: 0xD, binary 1101
E	11100	Data Bits, Value: 0xE, binary 1110
F	11101	Data Bits, Value: 0xF, binary 1111
S	11001	Set; Logical "on" or "true"
R	00111	Reset; Logical "off" or "false"
Q	00000	Quiet; Absence of activity on medium (e.g. broken line)
I	11111	Idle; Normal condition of the medium between transmissions
H	00100	Halt; Forced logical break in activity on the medium
T	01101	Terminate; Terminates all normal data transmission sequences
J	11000	Start Delimiter; First symbol in Starting Delimiter (SD) sequence
K	10001	Start Delimiter; Second symbol in SD sequence

FDDI - Clock variation

Frame is preserved, but space (IDLE) between frame can shrink or expand depending on variation in clock rates between nodes. This problem arises because FDDI is a **distributed** system implementation (there is not a single master, master clock etc.). Accommodating clock variation requires some careful design.



FDDI - Elasticity buffer and smoother function

Elasticity buffer:

Each point-to-point link is clocked by the Tx node. Stations can have different clock rates $\delta f < 0.01$ %. Insert IDLE preamble symbols. At least 16 IDLE before each frame. Elasticity buffer in other stations may change the length of the IDLE pattern. E-buffer is a FIFO which is filled half way before bits are removed. This requires pointer control in the FIFO.

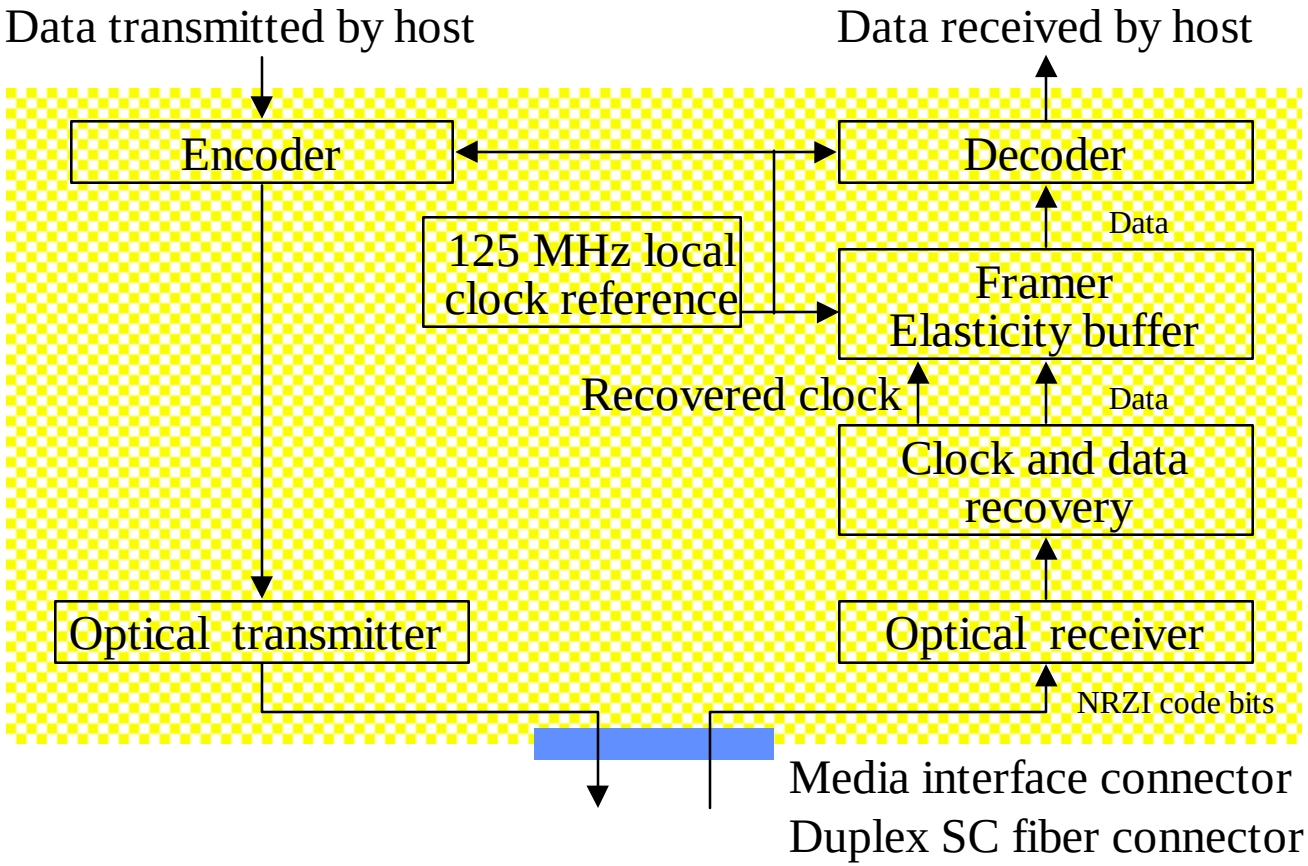
Smoother function:

Need to compensate for E-buffer deleting too many symbols from the same preamble. Unconstrained preamble shrinkage can result in loss of frames. Smoother function absorbs surplus symbols from longer preambles and redistributes them into shorter preambles. Smoother comes after E-buffer in each station.

Because this is a distributed algorithm for a statistical process it is very hard to test.

DEC implemented a 200 node ring in hardware to prove the smoother function.

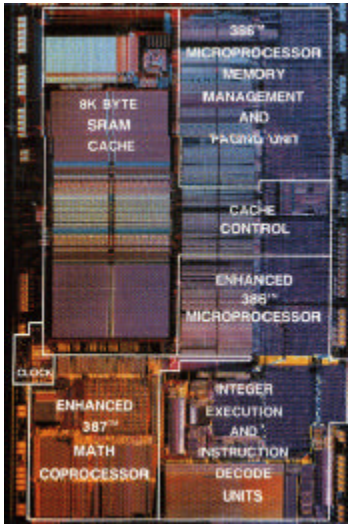
FDDI - Physical layer model of station node



The Impact of Integrated Circuits

1958 - Jack Kilby, hired by Texas Instruments form the University of Illinois, demonstrated the first integrated circuit. Patent applied for in 1959 and granted in 1964.

1970 - Gilbert Hyatt patents first integrated circuit computer.



1971 - Marcian E. Hoff, Federico Faggin and Stanley Mazor create the first commercial microprocessor, the Intel 4004.

1974 - Intel 8080 first 8-bit microprocessor.

1976 - Zilog launched the Z80 and created a large market with an inexpensive and faster version of the 8080.

1978 - Intel launched 8086 and is first commercially successful 16-bit microprocessor.

1979 - 68000 launched by Motorola.

1985 - Intel launched i386, a very successful 32-bit processor. 3 - 4 million instructions per second.

Microprocessors

1993 - Intel Pentium introduced in March. Clock speed 120 - 200 MHz.

1995 - Intel Pentium Pro introduced in November.

1997 - Intel Pentium MMX introduced in January. Clock speed 150 - 233 MHz.

1997 - AMD K6 MMX introduced in 2Q. Clock speed 233 - 300 MHz.

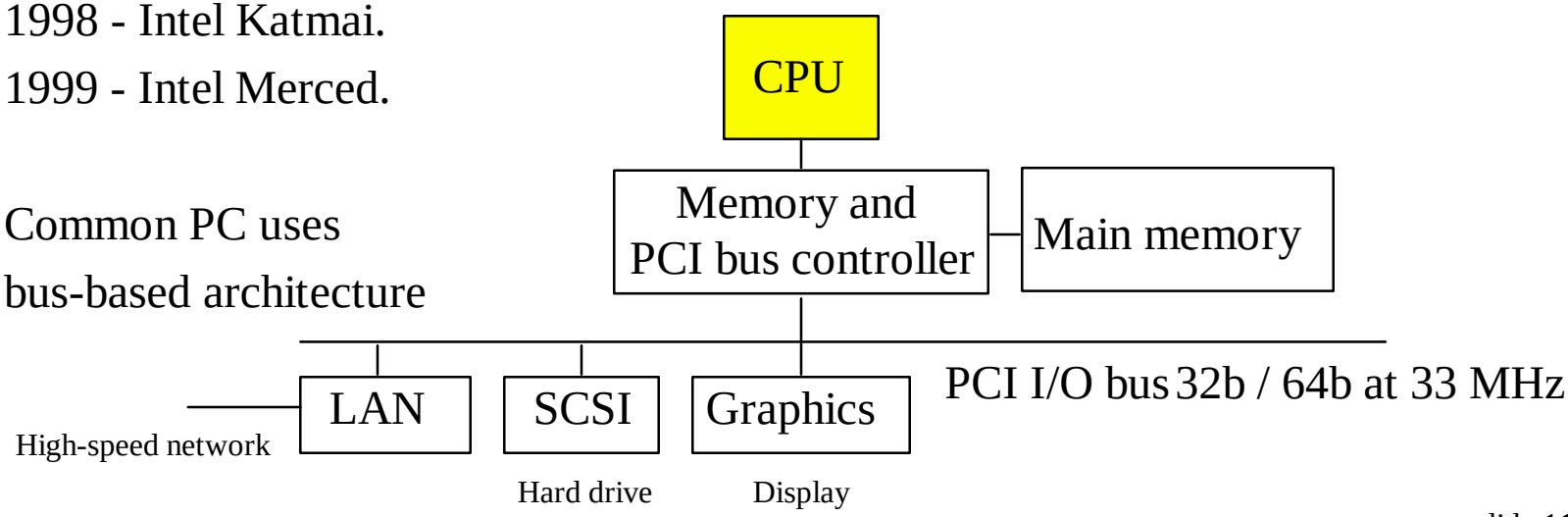
1997 - Intel Pentium II introduced in 2Q. Clock speed 233 - 300 MHz.

1997 - Intel Deschutes in 3Q. Clock speed 300 - 433 MHz.

1998 - Intel Katmai.

1999 - Intel Merced.

Common PC uses bus-based architecture



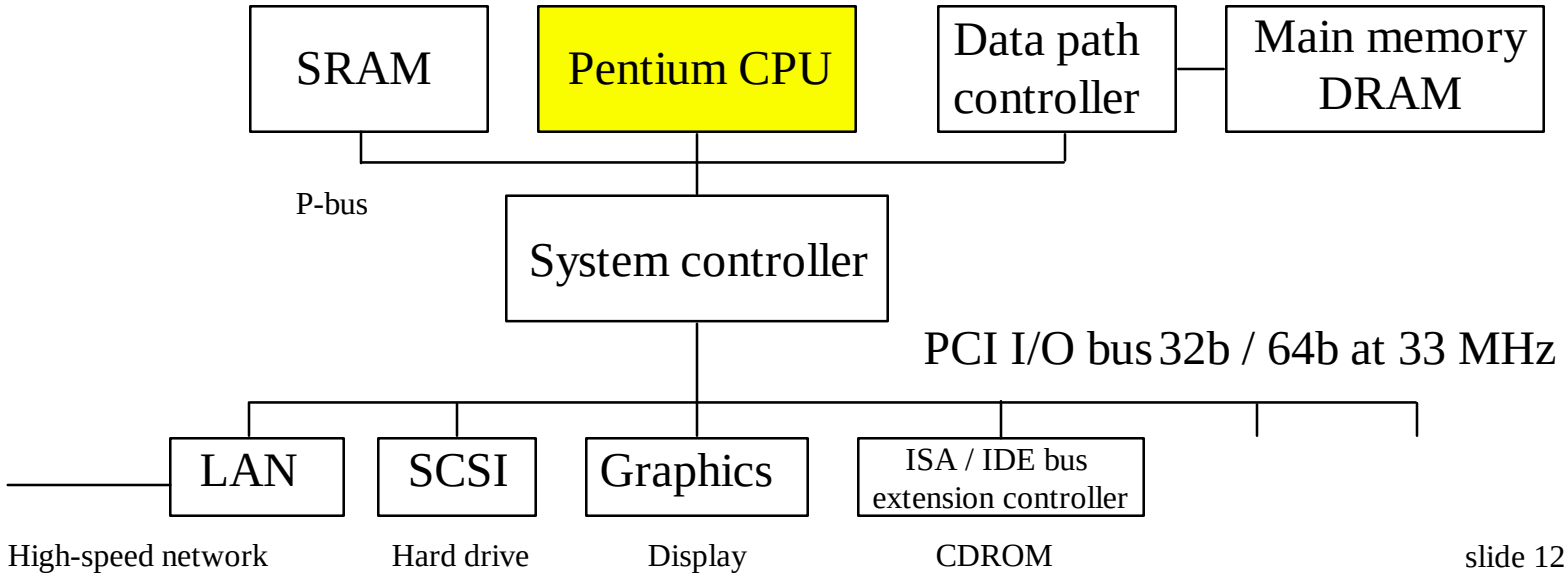
Growth of CPU - Network performance mismatch

On average CPU speed doubles every 18 months

Main memory data transfer speeds increase 10% every 18 months

Network node (end-system) operates at memory to I/O data transfer speeds

To obtain high performance need to minimize number of operations involving main memory



Hardware improvements reduce bottlenecks

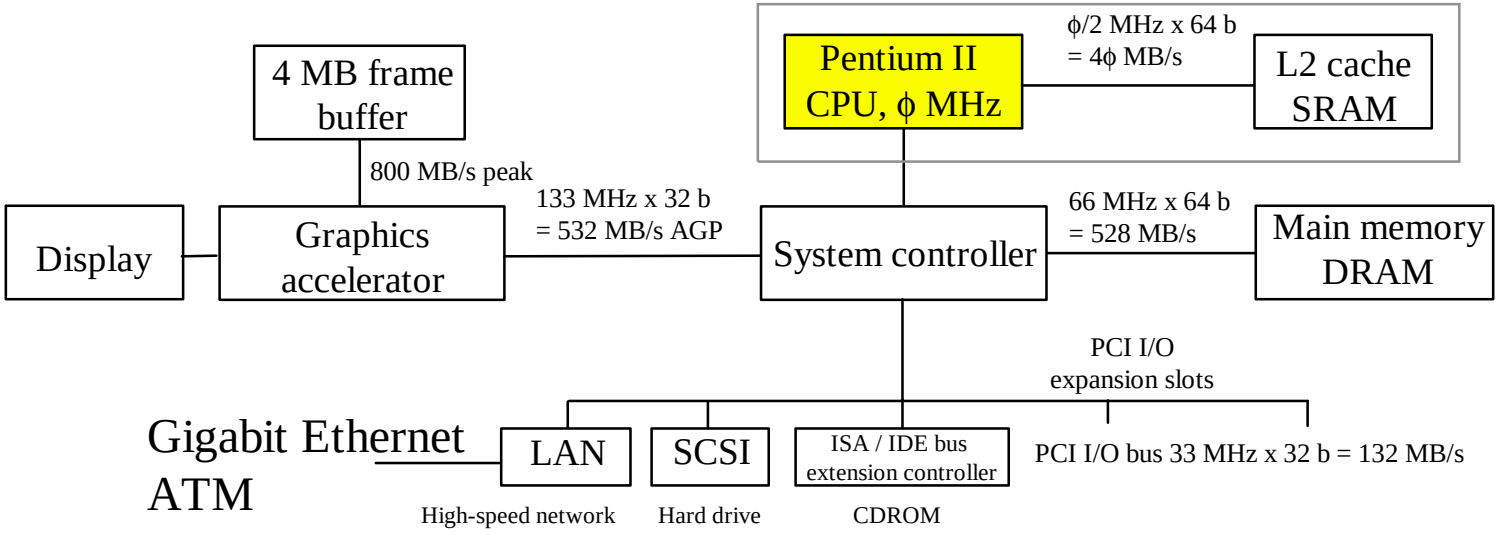
Intel approach is to increase number of dedicated buses (similar to SGI machines designed more than 5 years ago).

Intel Pentium II has a 32 bit-wide Accelerated Graphics Port (AGP) and dedicated non-blocking 64 bit-wide port for 512 kB SRAM L2 cache.

For more information on AGP see:
http://developer.intel.com/drg/mmx/AppNotes/agp.htm

1997 - Jan. Intel Pentium MMX (150 - 233 MHz)
1997 - 2Q AMD K6 MMX (233 - 300 MHz)
1997 - 2Q Intel Pentium II (233 - 300 MHz)
1997 - 3Q Intel Deschutes (300 - 433 MHz)
1998 - Intel Katmai
1999 - Intel Merced

Processor	*Clock (MHz)*	*SPECint95*	*SPECfp95*
Pentium II	266	10.8	6.9
DEC Alpha	266	7.9	11.8
Pentium II	300	11.6	7.2
DEC Alpha	333	9.8	12.5
DEC Alpha	500	15.0	20.4



Network Protocols

- The need for protocols
- How protocols work
  - physical layer and Media Access Control
  - network and transport layer
    - TCP/IP
    - UDP/IP
- Impact of protocols on high bandwidth multimedia networks

The OSI -ISO Network Model

Open Systems Interconnection - International Standards Organization

7. Application	Common protocols such as network virtual terminal, file transfer protocol (FTP) electronic mail, and directory lookup.
6. Presentation	Encoding/decoding including compression and cryptography.
5. Session	Communication between processes including data exchange, remote Procedure Call (RPC), synchronization, and activity management.
4. Transport	Lowest level at which messages are handled. Segmentation and reassembly of data to and from session layer. Transmission Control Protocol (TCP), Internet Protocol (IP), User Datagram Protocol (UDP)
3. Network	Flow control to avoid congestion and also customer use and accounting. Link Layer Control (LLC).
2. Data Link	Presentation of error-free transmission to the network layer. Creates data frames and receives acknowledge frames. Media Access Control (MAC)
1. Physical	Physical Layer Protocol (PHY) specifies coding (4B/5B), clock synchronization. Physical Medium Dependent (PMD) sublayer provides digital baseband between nodes. This layer specifies fiber-optic drivers, receivers, mechanical, cables, connectors, optical signal requirements including power levels, jitter and BER.

TCP/IP versus OSI - ISO reference model

1988 National Institute for Standard Technology (NIST) mandated ISO-complient protocols for all computers sold to the US Government.

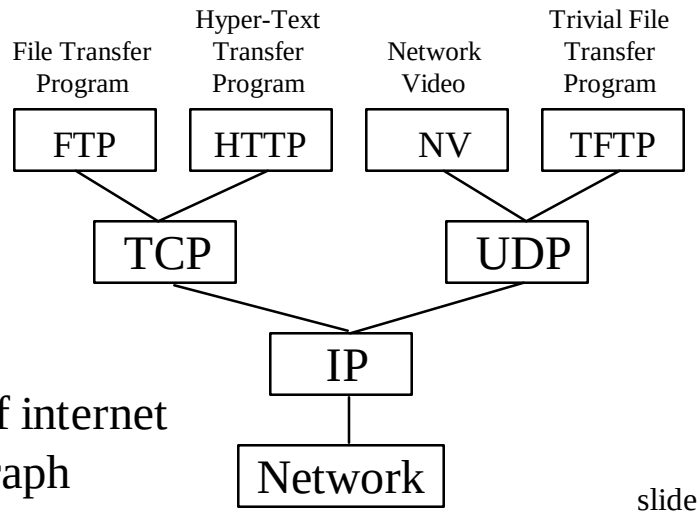
TCP/IP still dominated so edict was rescinded in September, 1994.

UCB UNIX distribution in 1980's included the TCP/IP protocol suite.

It works.

It is available.

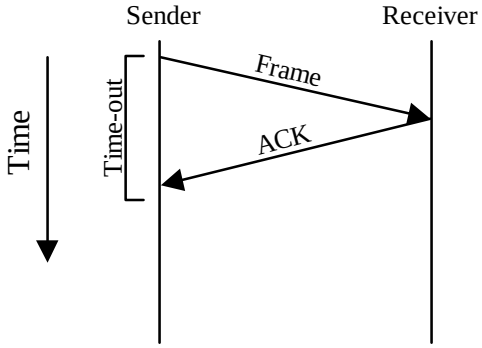
Don't fix what doesn't need fixing.



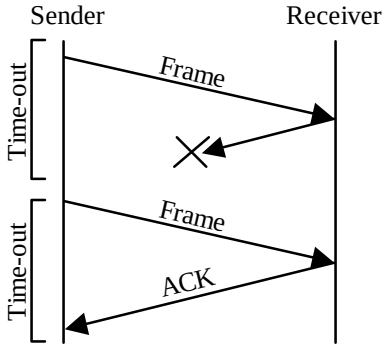
Example of internet protocol graph

Automatic Repeat Algorithm

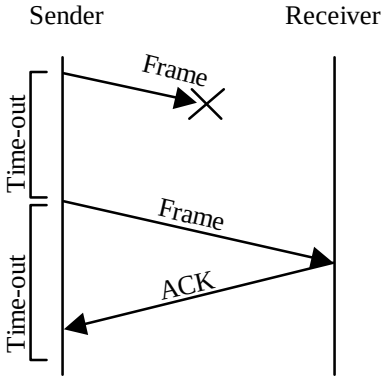
Stop and wait (time-out) ARQ is not an efficient way to use network bandwidth



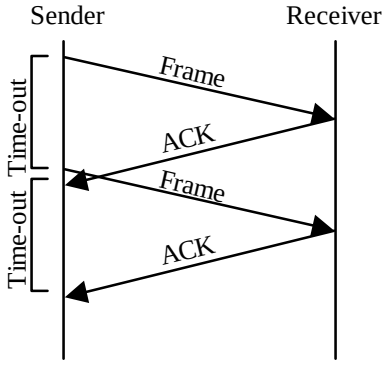
ACK received before timer expires



The ACK is lost



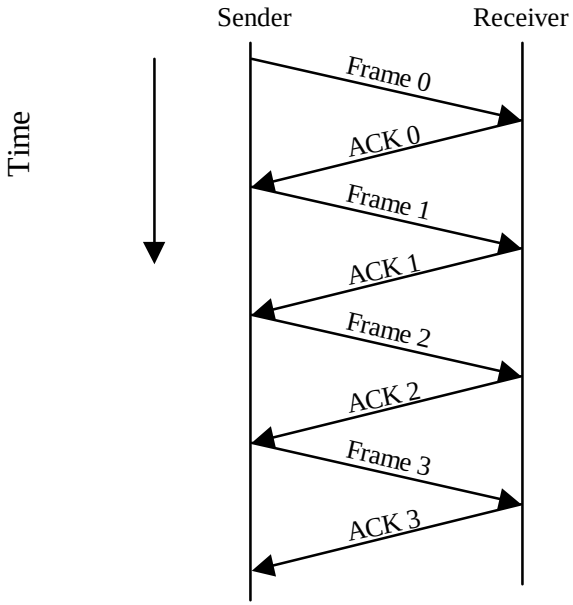
Original frame is lost



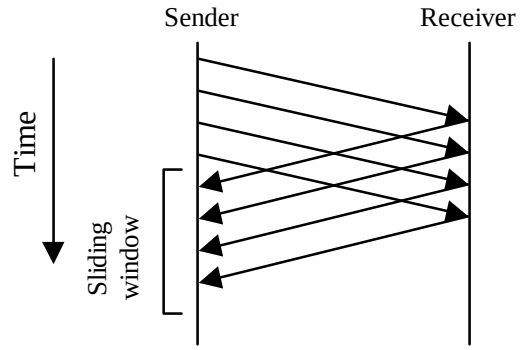
Time-out occurs too soon

Sliding Window

Sliding window uses sequence number and sliding window to more efficiently use network bandwidth



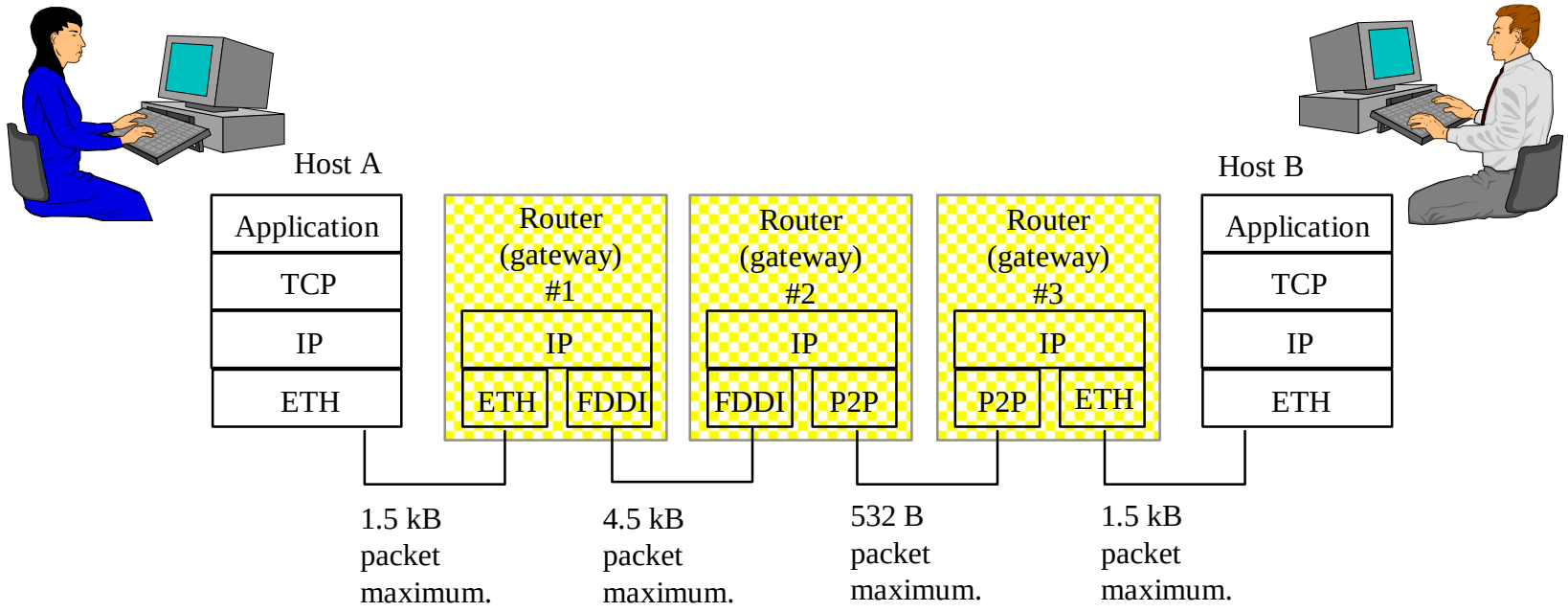
Stop and wait with one-bit sequence number



Sliding window

Internetworking (the creation of a network of networks)

Two important problems to address are *heterogeneity* and *scale*



Internet Protocol (IP) is the key tool used to build scalable heterogeneous internetworks

The IP datagram

Carries enough information to let the network forward packet to destination in a connectionless way.

Unreliable or best effort service - network does not attempt recovery from failed delivery.

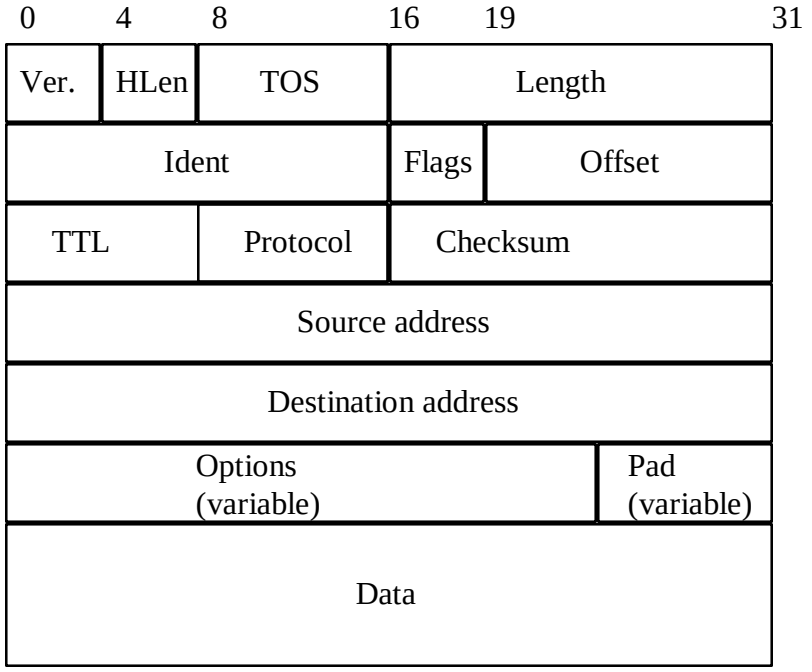
IP Packet format (IPv4)

Header is arranged in 32-bit words

Ver.	Current version is IPv4
HLen.	Length of header in 32-bit words
TOS	Type of service (low delay, high-throughput, or high reliability)
Length	16-bit length of datagram (including header) in Bytes. Maximum size is 65,535 Bytes

Because the physical network may not support datagrams this long IP supports fragmentation and reassembly
TTL time to live or hop counter (64 hop default)

Protocol	Demultiplexing key, e.g. TCP = 6 , UDP = 17
Checksum	16-bit checksum of the IP header
Options	Rarely used



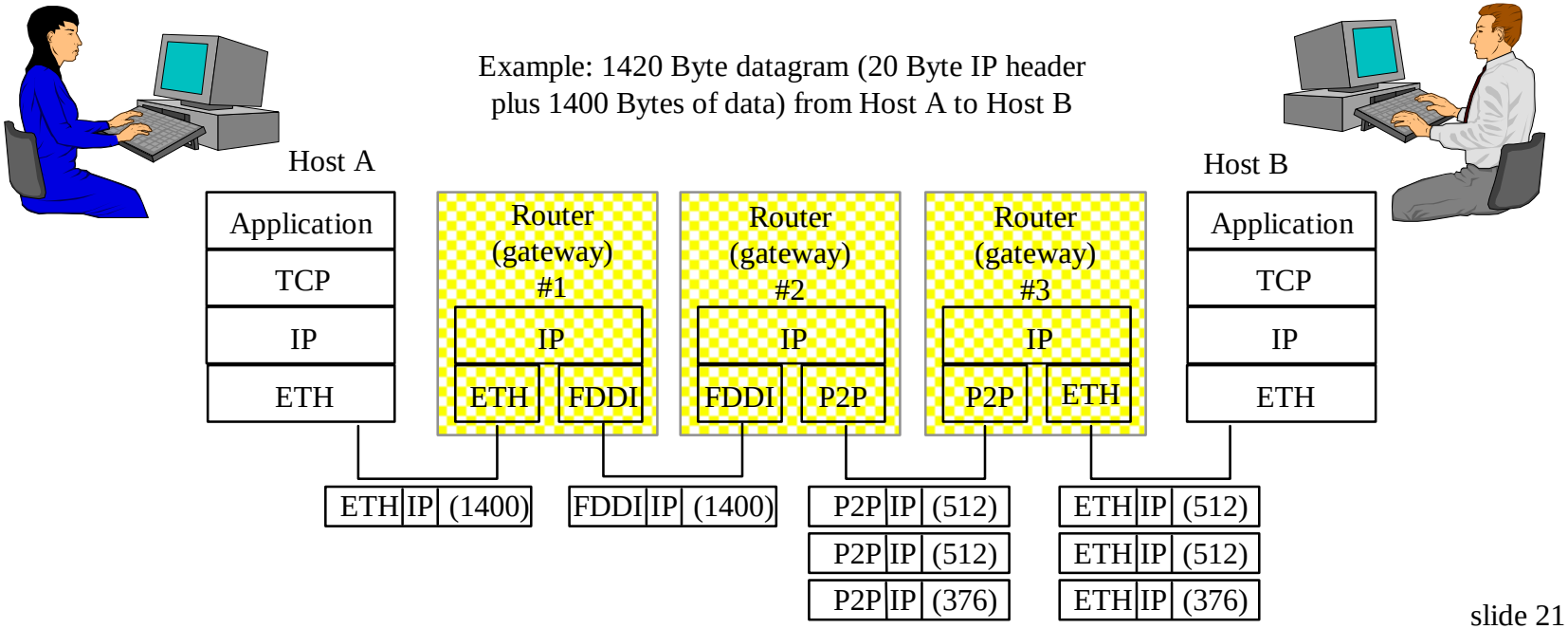
Fragmentation and Reassembly

Need to accomodate different physical network maximumn packet sizes

e.g. Ethernet 1.5 kB and FDDI 4.5 kB.

Every network type has a Maximum Transmission Unit (MTU) which is the largest IP datagram that it can carry in a frame.

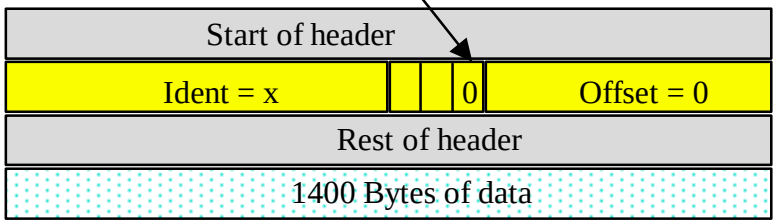
If MTU decreases over part of a network, it may be necessary to fragment the packet.



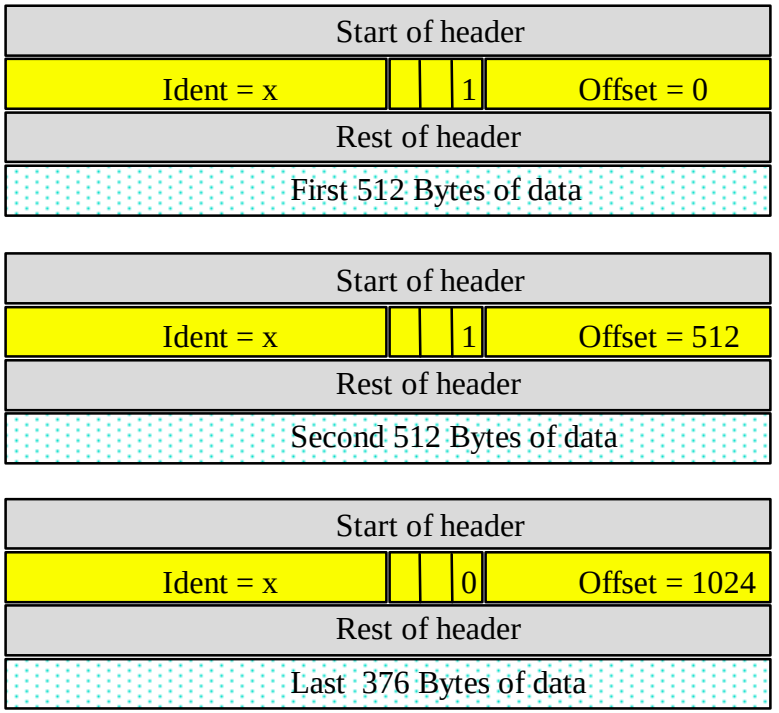
Fragmentation and Reassembly header format

Unfragmented packet

M-bit set to zero indicating that there are no fragments to follow



Example of fragmented packet
(implementation is an exercise in bookeeping!)



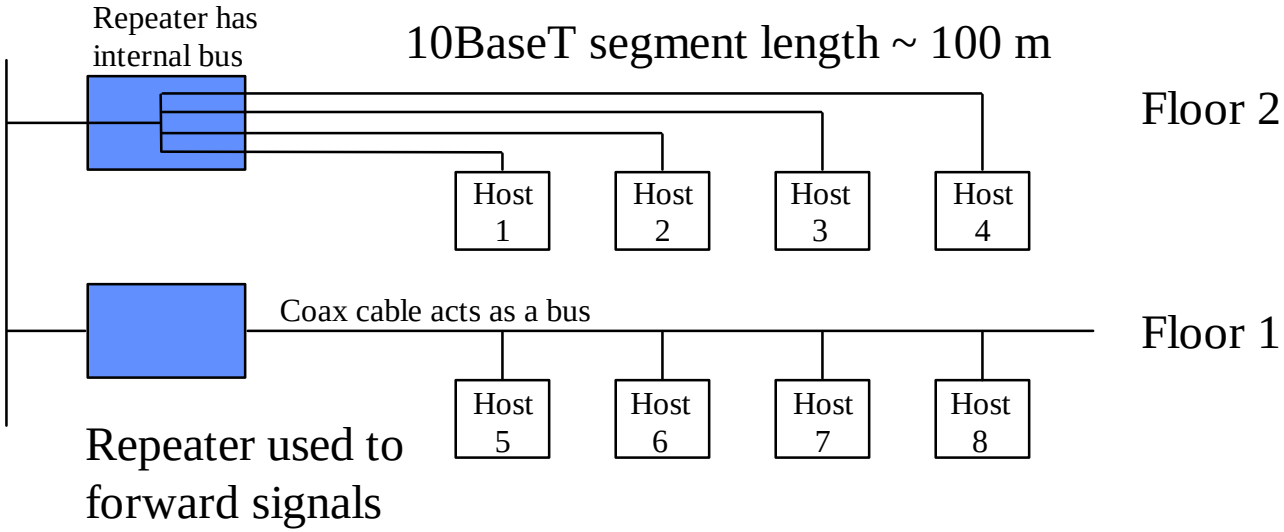
Note: The offset field counts on 8-Byte units of data

CSMA/CD Ethernet

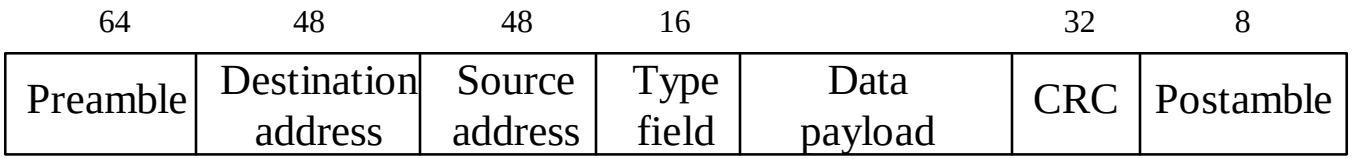
Local Area Network (LAN) developed at Xerox Palo Alto Research Center (PARC) in mid-1970s became IEEE 802.3 standard. 10 Mb/s data rate over shared medium, 1024 maximum number of nodes, 1500 m maximum length. Uses Carrier Sense, Multiple Access with Collision Detect (CSMA/CD).

Carrier Sense - all nodes can distinguish between idle and busy link

Collision Detect - node listens for interference from other nodes as is transmits



Ethernet Frame Format



64 bit preamble is sequence of alternating 1 and 0 for receiver synchronization with signal.

Every Ethernet adapter attached to a host has a unique 6-Byte address e.g. 8:0:2b:e4:b1:2 is 0001000:00000000:00101011:11100100:10110001:00000010

Ethernet standard defined by Xerox, DEC and Intel in 1978 uses 16 bit type field for demultiplexing to frame to higher level protocols. IEEE 802.3 standard uses this field to determine how long the frame is.

Maximum data payload is 1500 Byte.

Cyclic Redundancy Code (CRC-32) is used for error checking.

Postamble indicates end of frame.

Ethernet Receiver

For an Ethernet segment the sending adapter attaches preamble, CRC and postamble and an accepting receiving adapter removes them.

An Ethernet adapter receives all frames. The adapter only accepts and passes to its host:

Unicast mode - Frame with its destination address.

Broadcast mode - Frame with broadcast destination address consisting of all 1s or (ff:ff:ff:ff:ff:ff).

Multicast mode - Frame with multicast address (first bit 1) it has been programmed to accept.

Promiscuous mode - adapter programmed to pass all frames to host.

Ethernet Transmitter Algorithm

If adapter has a frame to send and line is idle it transmits frame immediately (connectionless - no negotiation with other adapters).

Fairness of access mechanism: Adapter must wait at least 51.2 µs before transmitting another frame to allow others access to network.

If two or more adapters begin transmitting at the same time Ethernet detects a frame collision. The adapters must transmit a minimum 512 bit (64 Byte) frame before aborting the transmission. The guaranteed minimum 51.2 µs jam time ensures that the collision can be detected over the maximum network length of 1500 m (note maximum round-trip time-of-flight of signal in cable is only 15 µs).

After collision, the adapter doubles the maximum wait time before retrying (exponential backoff). The number of times the maximum wait time is doubled is limited to 10.

Adapter will report an error to the host if after 16 tries it still cannot transmit the frame.

Why is Ethernet Successful?

Ethernet is inexpensive and easy to administer. It is easy to increase or decrease the number of nodes in the network.

In practice Ethernets are divided into subnetted segments which have less than 254 nodes and typically ~5 μs network delay. This is much less than the standard allows (maximum of 1024 nodes and 51.2 μs maximum delay) and helps keep the network data traffic load low.

Because network capacity is wasted by collisions, Ethernet works best under lightly loaded (< 30 % capacity) network traffic conditions.

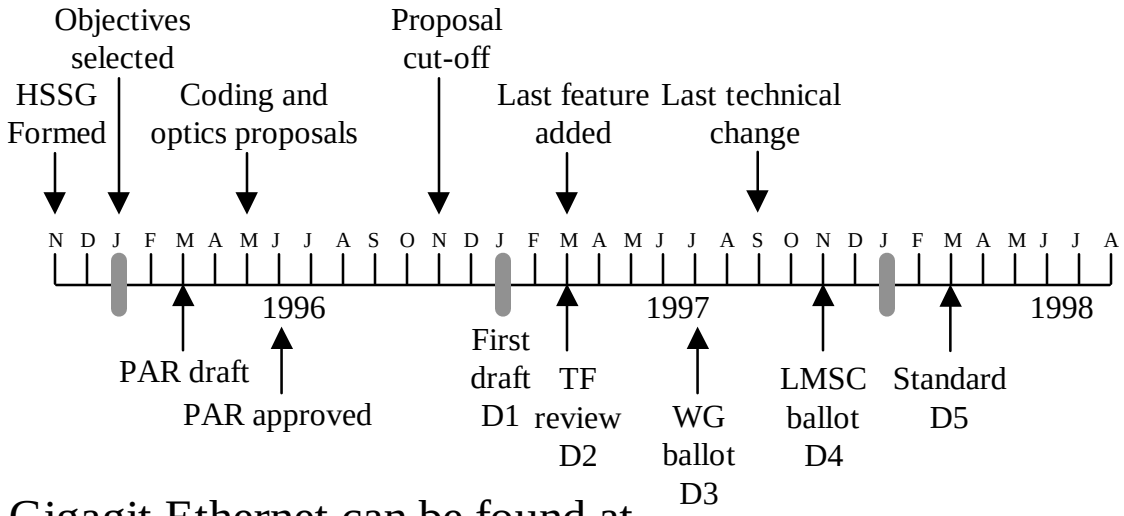
Ethernet does not provide link-level flow control. End-to-end flow control is provided by the host. This avoids instabilities such as hosts continuously pumping frames into the network.

Case Study: Gigabit Ethernet

- The new IEEE 802.3 standard in high bandwidth networks
  - why Gigabit Ethernet
  - how it works
  - cost compared to ATM
  - when and where you will see it
  - impact on multimedia and what to expect beyond Gigabit Ethernet

IEEE 802.3z - The Gigabit Ethernet Standard

Time-line of Gigabit Ethernet Task Force



General information on Gigagit Ethernet can be found at

http://www.gigabit-ethernet.org/

Drafts in pdf of standard developed by the Gigabit Task Force can be found at

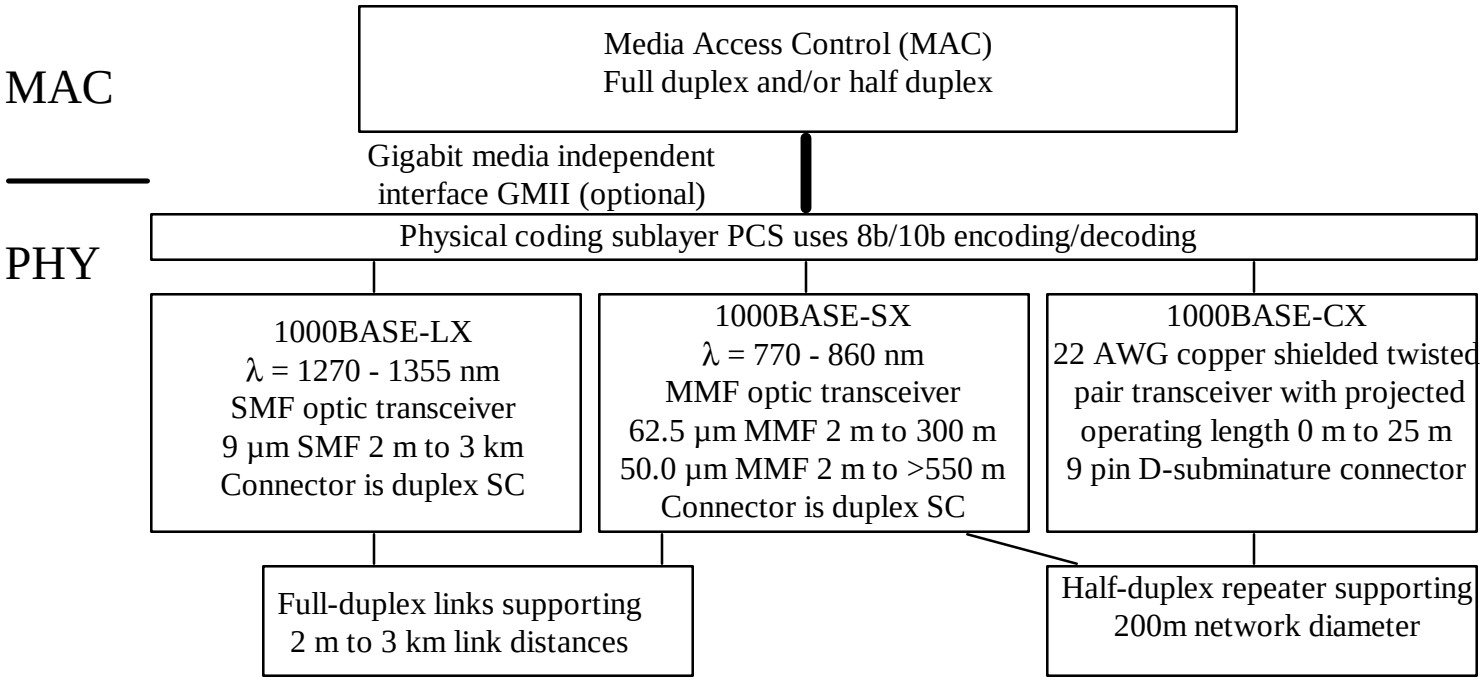
http://grouper.ieee.org/groups/802/3/z/private

user: 802.3z

password: go_fastR

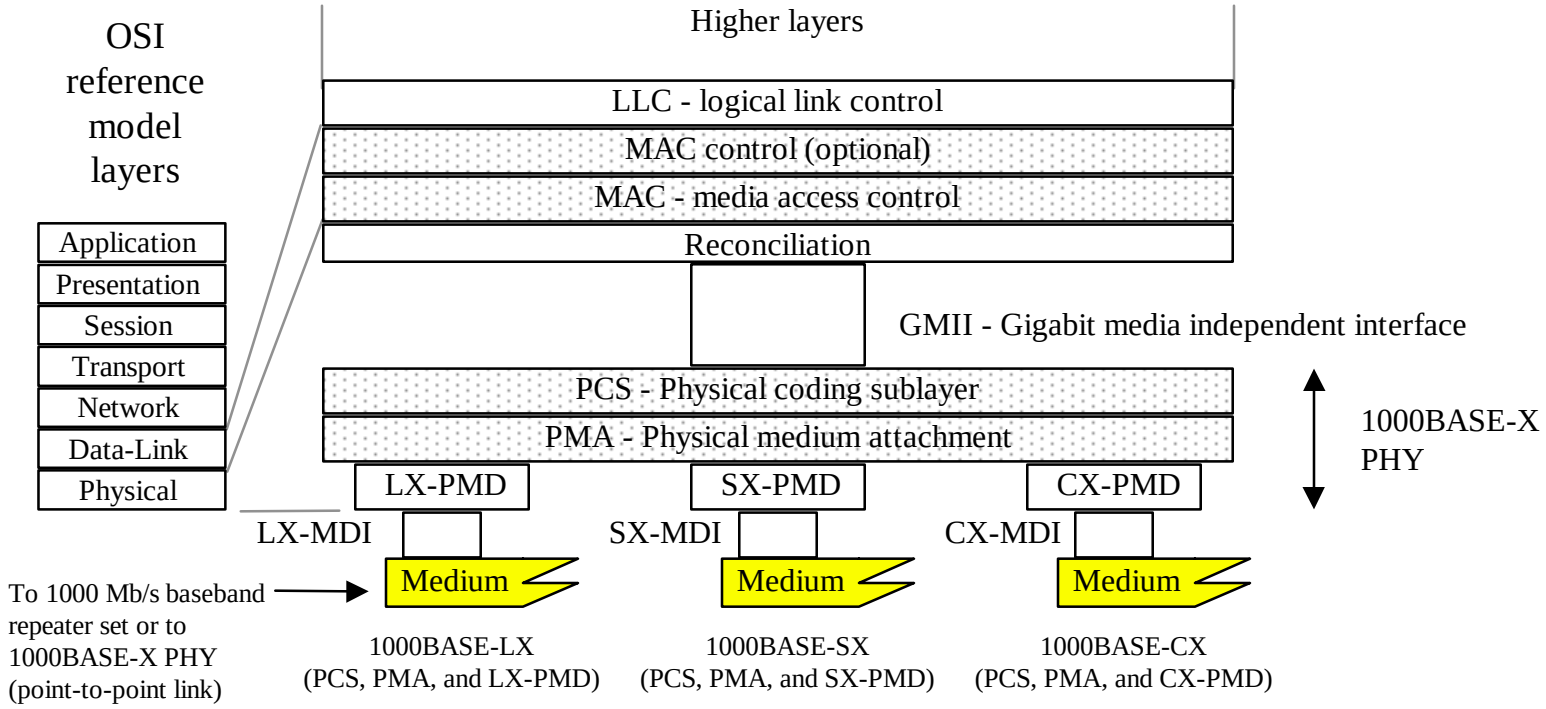
Functional elements of Gigabit Ethernet technology

Enhanced version of Fiber channel (ANSI X3T11) FC0 physical and signaling interface (ANSI X3.230-1994). Includes 8b/10b coding and signaling rate increased to 1.250 GBd to achieve a 1 Gb/s data rate.



1000BASE-X

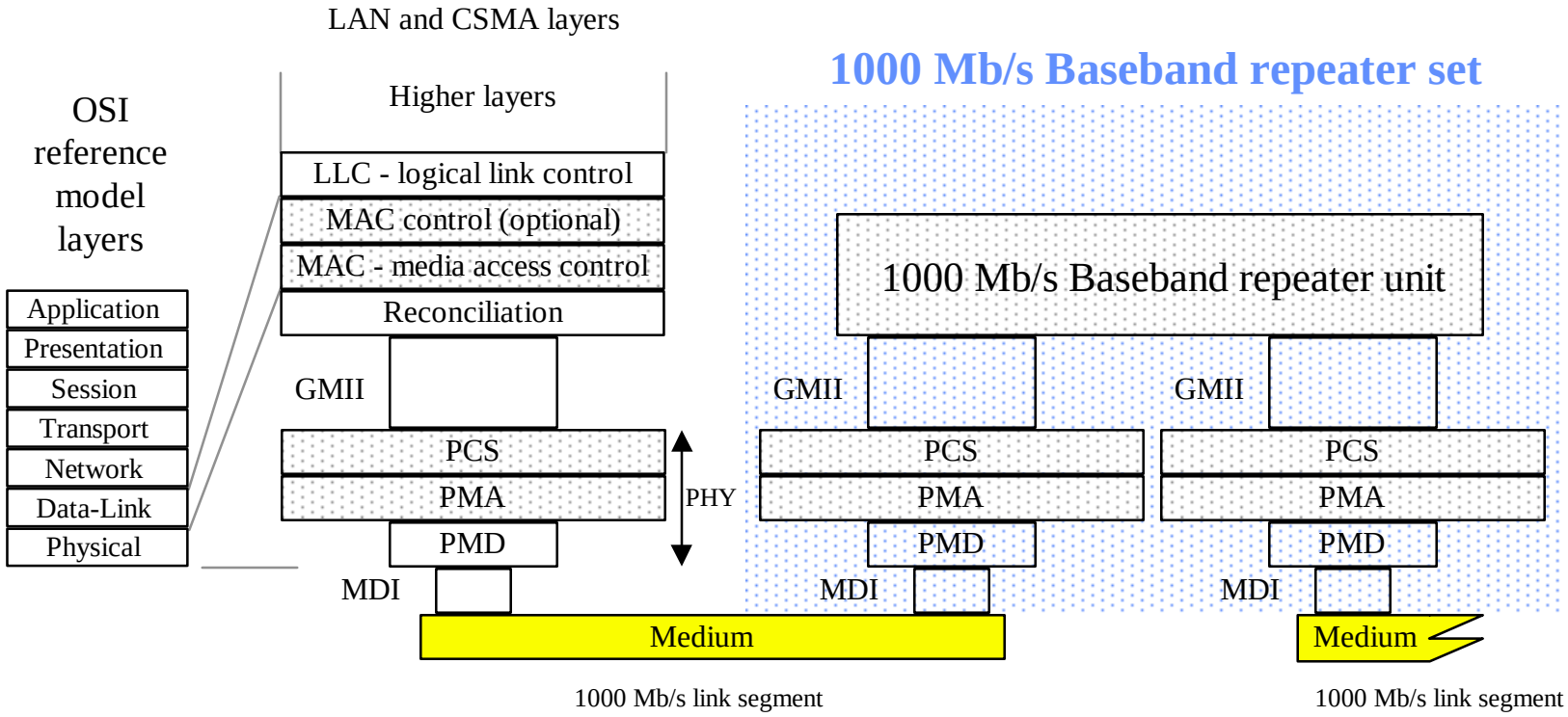
LAN and CSMA layers



MDI - Media dependent interface
PHY - Physical layer device
PMA - Physical medium attachment
PMD - Physical medium dependent

LX-PMD = PMD for fiber - long wavelength
SX-PMD = PMD for fiber - short wavelength
CX-PMD = PMD for fiber - 150Ω balanced copper cabling

1000 Mb/s repeater set



MDI - Media dependent interface
PMS - Pysical coding sublayer
PHY - Physical layer device
PMA - Physical medium attachment
PMD - Physical medium dependent

GMII - Gigabit media independent interface
LX-PMD = PMD for fiber - long wavelength
SX-PMD = PMD for fiber - short wavelength
CX-PMD = PMD for fiber - 150Ω balanced copper cabling

Parameterized Values

	Ethernet 10BaseT	GigabitEthernet
slotTime	512 bit	4096 bit
interFrameGap	9.6 µs	0.096 µs
attemptLimit	16	16
backoffLimit	10	10
jamSize	32 bit	32 bit
maxFrameSize	1518 Byte	1518 Byte
minFrameSize	512 bit (64 Byte)	512 bit (64 Byte)
extended Size	0 bit	3584 bits (448 Byte)
burstLength	12000 bit	12000 bit

Carrier extension concept

To preserve the 200 m network diameter of 100Base-T using half-duplex links, the 802.3z working group for Gigabit Ethernet implemented a techniques called carrier extension.

Whenever the Gigabit Ethernet adapter transmits a frame less than 512 Bytes the carrier extension mechanism maintains the total frame size at 512 Bytes by adding non-data carrier extension Bytes. If Gigabit Ethernet detects a collision during this period it sends out a jam signal so offending stations back-off and try again.

The 512 Byte limit (instead of 640 Byte) is possible because the number of repeater hops is reduced from two (100Base-T) to one (Gigabit Ethernet) and the built-in engineering safety margin (extra time) is eliminated.

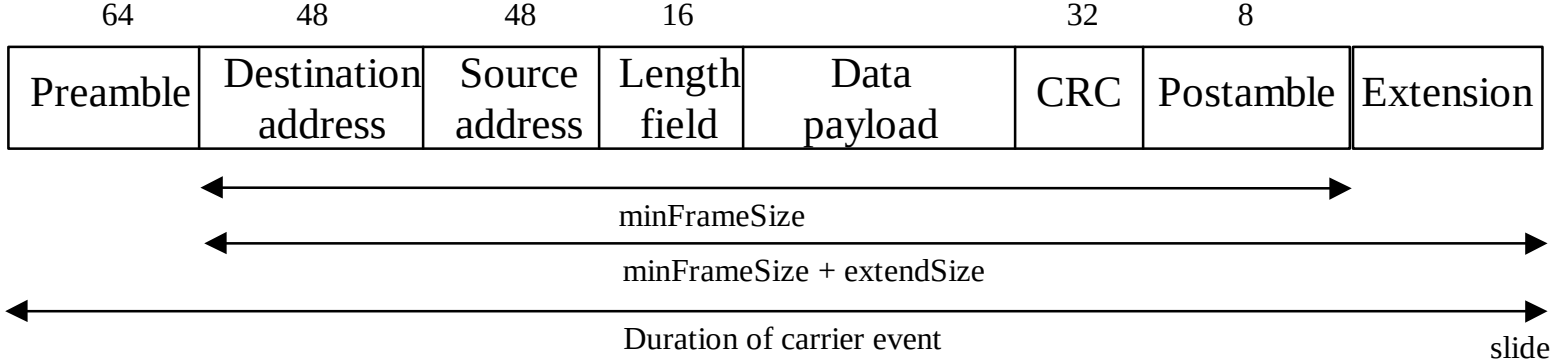
Recognizing the potential impact on sustained throughput, the 802.3z working group also implemented a frame bursting mode in which the transmitting station may sequentially transmit frames without contending for the medium for each frame for a period up to the burstLength.

Carrier extension impact on data throughput

In half-duplex mode operating at data rates above 100 Mb/s the slotTime used at lower rates is inadequate to accommodate network topologies of the desired physical size (200 m). Carrier extension appends non-data extension bits to frames that are less than minFrameSize + extendSize bits in length to give a transmission at least slotTime in size.

This approach results in reduced sustained throughput for small packets. Worst case, if traffic consists of only 64-Byte frames effective throughput is reduced to 120 Mb/s. At present, average Ehernet frame size is 200 - 500 Byte and Gigabit Ethernet would deliver only 300 to 400 Mb/s throughput.

None of the above is an issue for fiber-based duplex links.

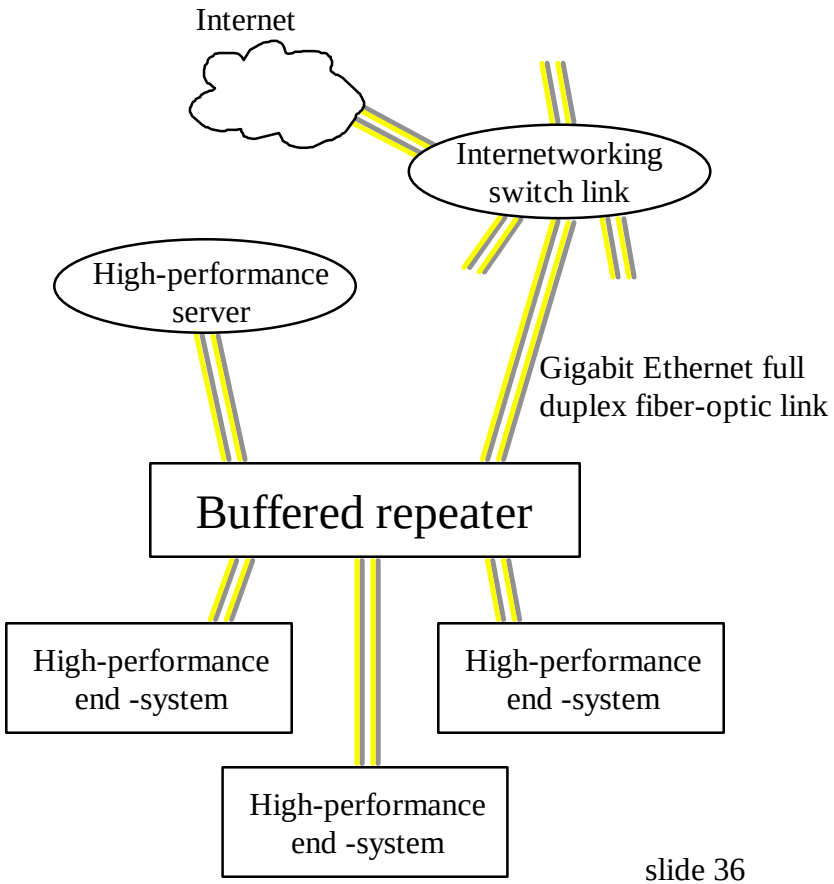


Gigabit Ethernet Buffered Repeater Full Duplex Implementation

The buffered repeater is a cost-effective way to maximize Gigabit Ethernet throughput.

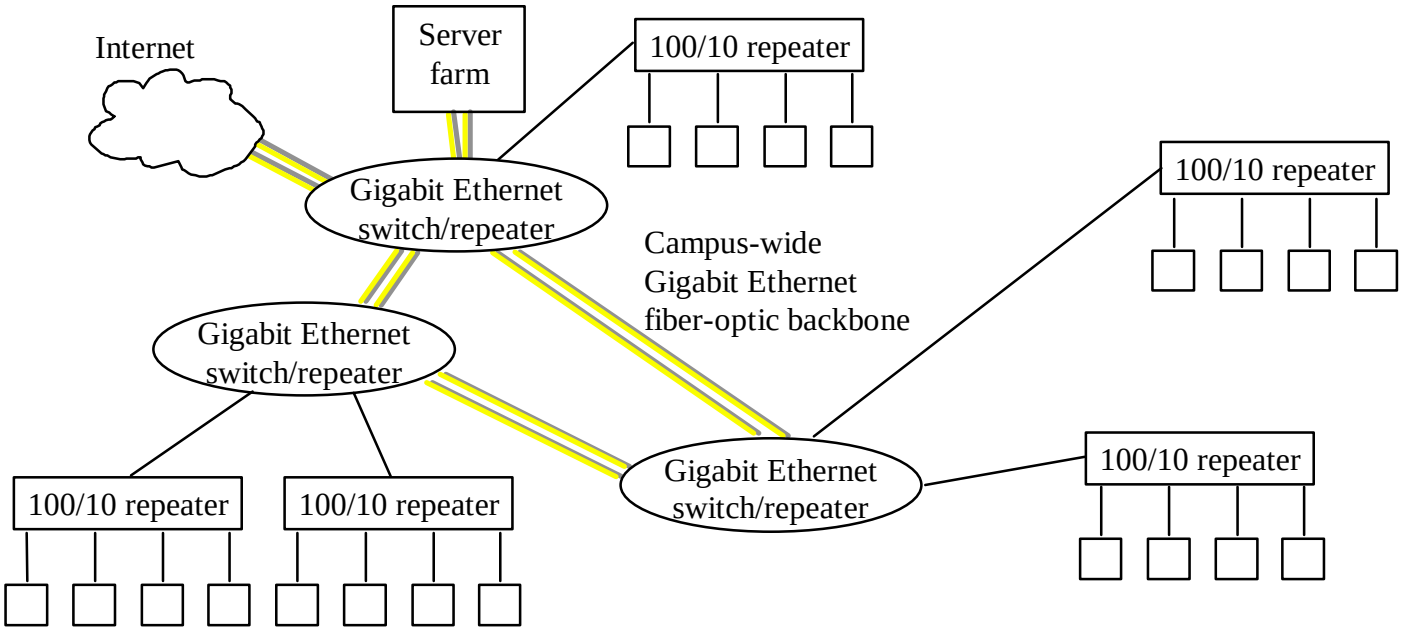
Buffered distributor/repeater uses full duplex links (a natural consequence of using fiber-optic physical layer).

Buffered repeater transmits each packet to all other connected nodes and can simultaneously receive on multiple ports by storing frames in its local memory. To avoid memory overflow there is flow control messaging to the transmitting node to stop sending while the repeater empties its buffers.



Gigabit Ethernet Backbone Implementation

Gigabit Ethernet used as a backbone linking many 10BaseT/100BaseT hubs and switches on a campus.



Quality of Service

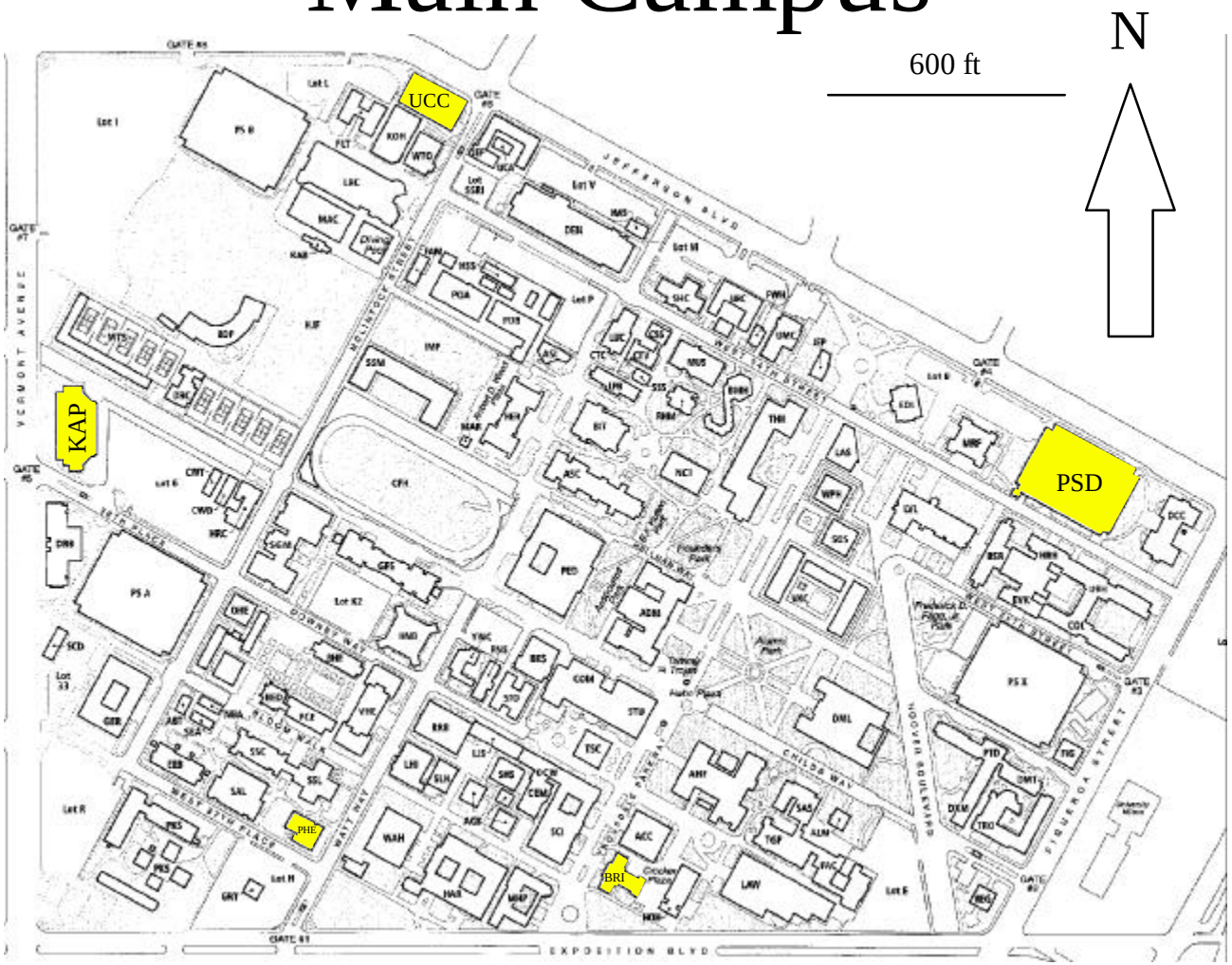
Ethernet volume will drive price low. This and compatibility with legacy Ethernet networks will be the key to its success compared to alternative network solutions such as ATM.

Network size is not an issue with full duplex point-to-point links implemented using fiber channel. Buffered distributor/repeater will deal with collisions and, to the extent that the buffers can handle, there will be fewer collisions.

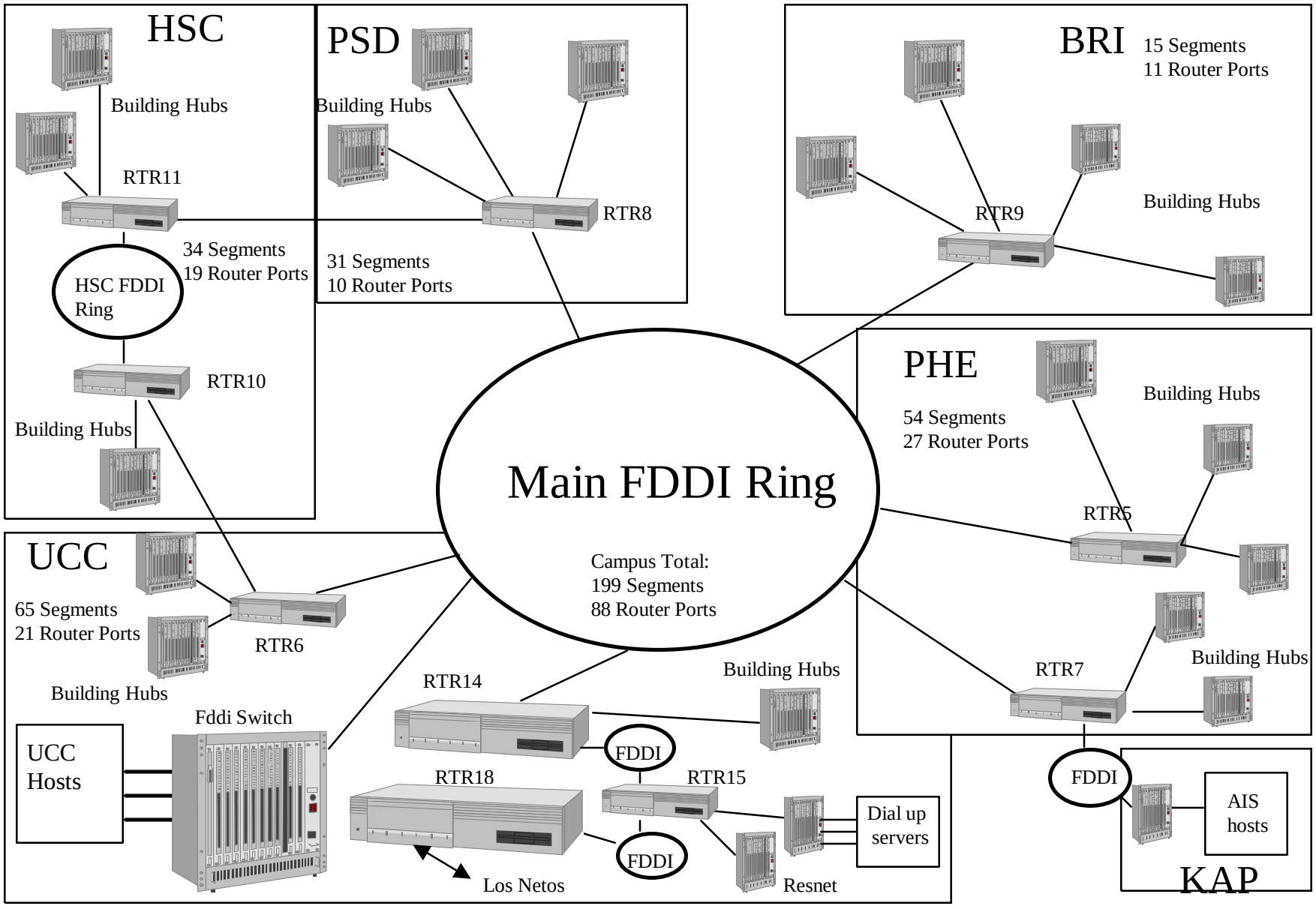
Guaranteed quality of service will always be an issue. However, in practical situations, just as with present Ethernet, FDDI, etc., network traffic load will be kept low by keeping the number of nodes per subnet low.

Support for real-time services cannot be entirely addressed by protocols. The physical layer of a network that provides guaranteed quality of service should offer connection admission control and predictable packet arrival. Gigabit Ethernet is a connectionless technology that transmits variable length frames and so cannot guarantee that time sensitive packets get priority.

University of Southern California
Main Campus



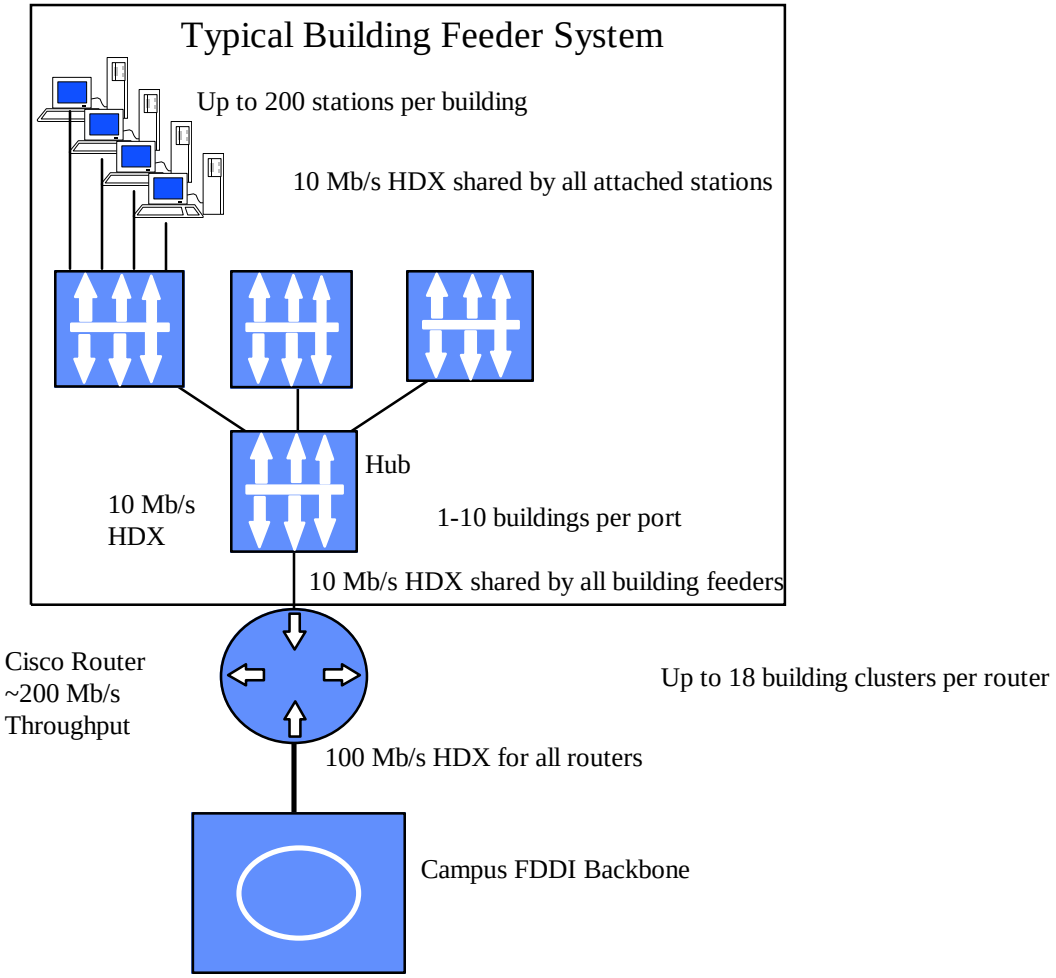
Current USC Production Network Configuration



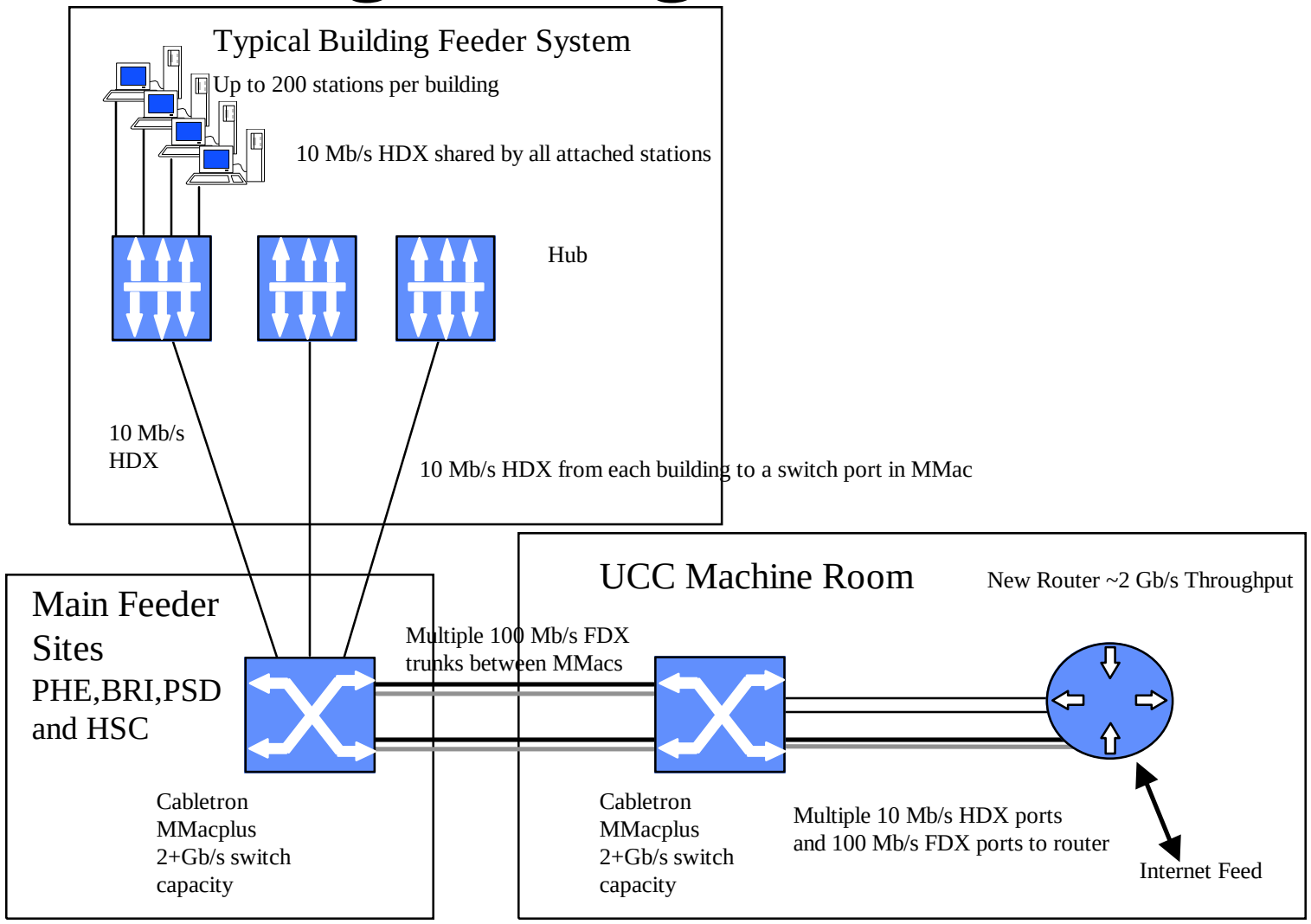
Evolving USC Network Usage

- Networks originally based on shared 10 Mb/s segments
  - Server for clients were local.
  - Programs were stored on local client not server.
  - 80-90% of data accessed was on local server.
- What has changed
  - Web has dense graphic content.
  - More data referenced is off the Web or “Enterprise” servers.
  - New desktop machines can handle streams over 10 Mb/s using GUIs increase bandwidth needs.
  - New applications such as data sharing, video and sound are driving up needs.
  - A much larger percentage of users in a building are using network at same time.
  - Usage pattern has evolved to at best 50% of data is from server(s) on local segment.
  - Massive data intensive resources are appearing on the Internet.

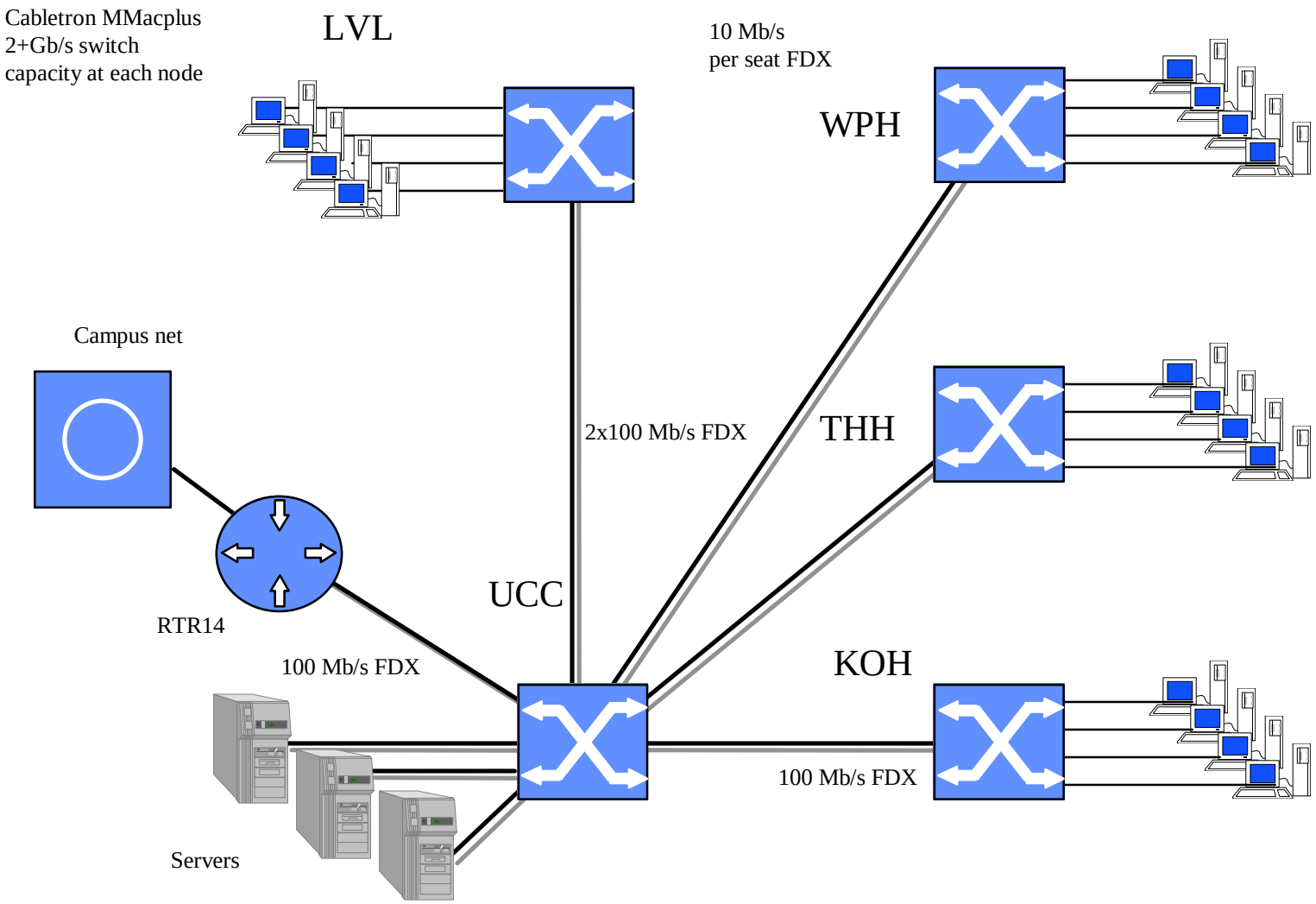
Current USC Router Configuration



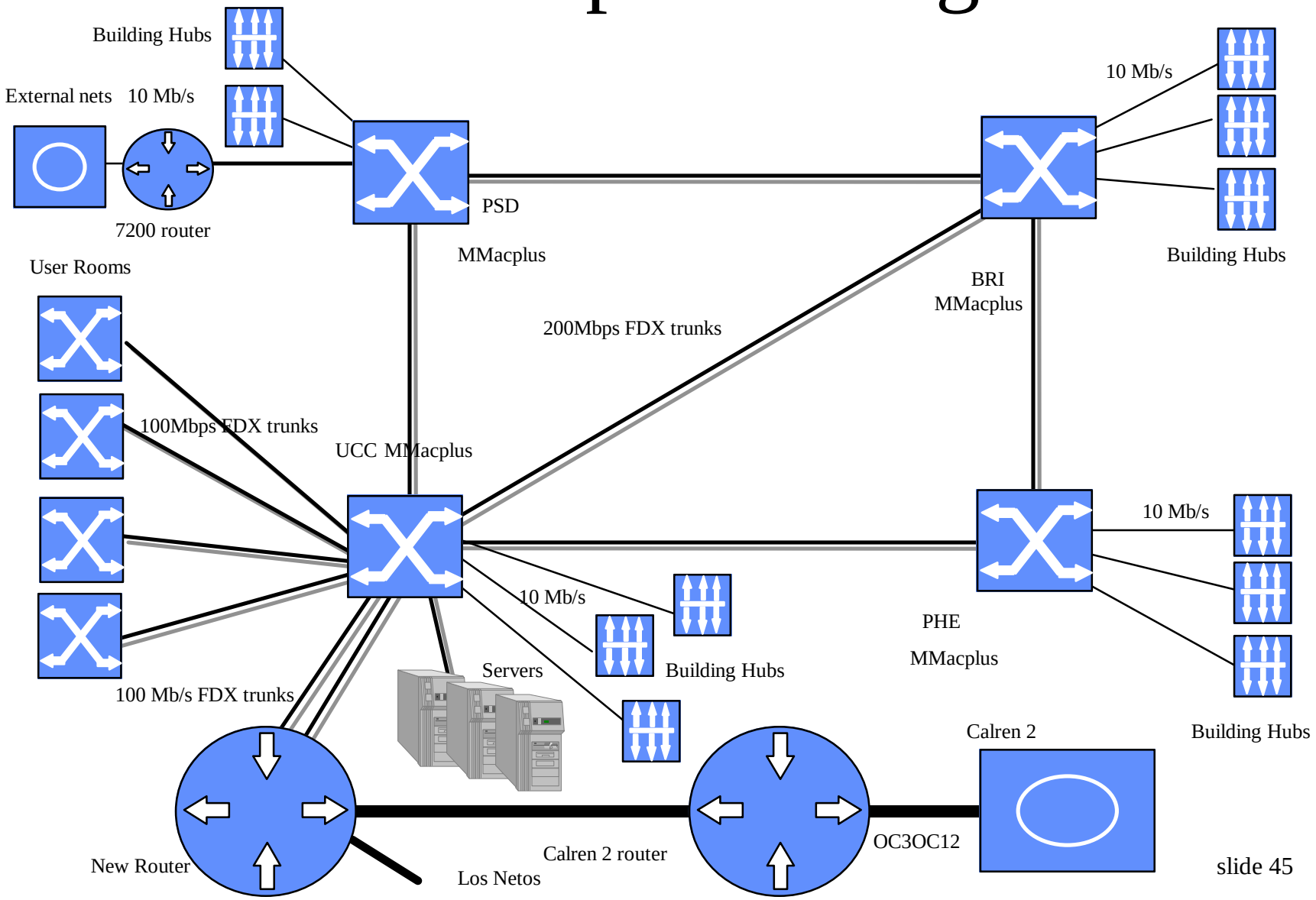
USC Building Configuration Phase 1



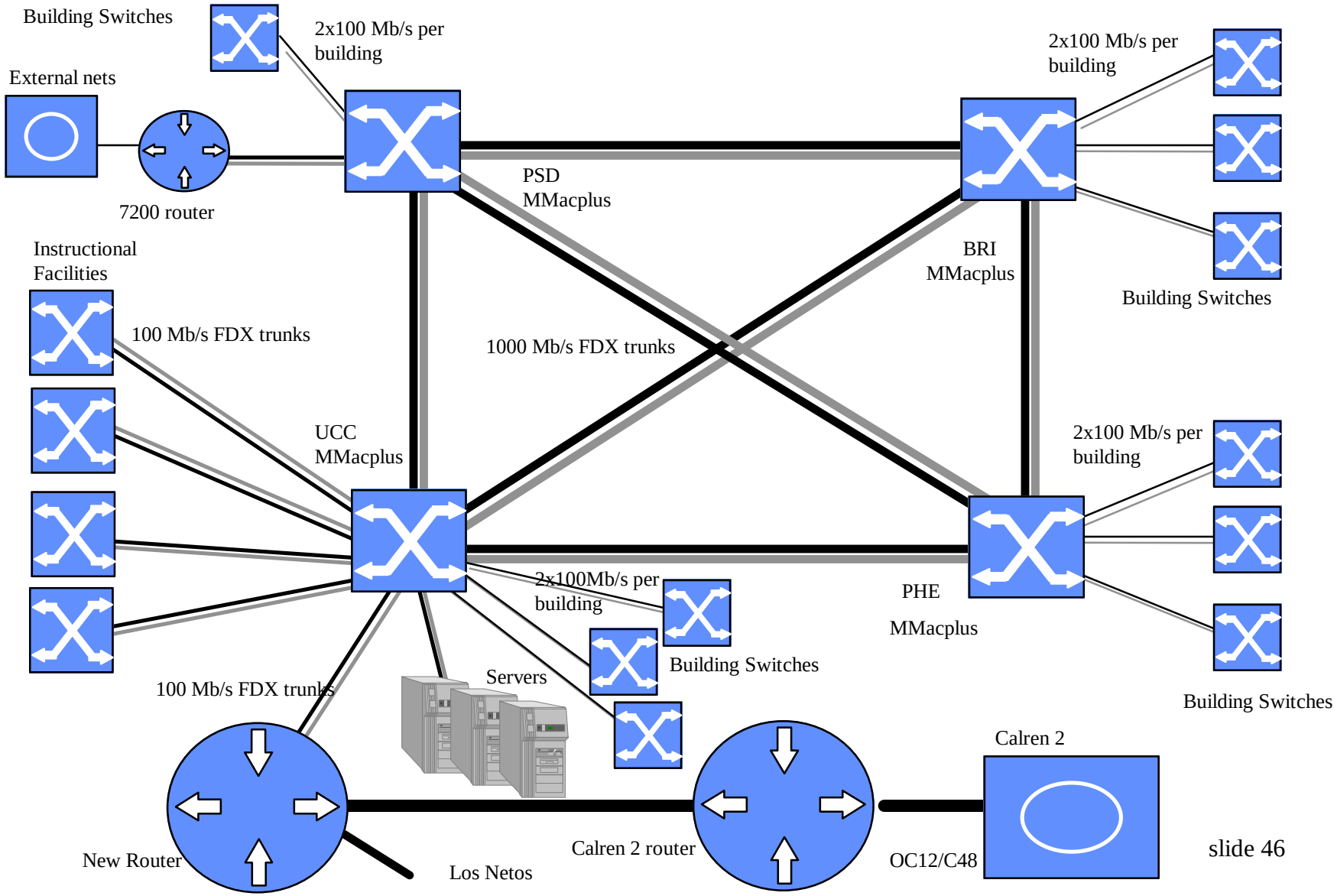
USC Public Facilities Phase 1



Phase 1 Campus Configuration



USC Phase 2 Configuration



USC Phase 1 Plan (Summer 1997)

- Replace collector routers with collector switch hubs.
- Put redundant Cisco 7500 at UCC to collect nets from collector PHE, UCC, BRI and PSD switch hubs (UCS acquires PSD, UCC and PHE hubs in place).
- Install switch hub at BRI (Business has already acquired this month).
- Install switch hub at PSD (UCS acquire).
- Install Cisco 7200 in PSD to collect “off campus” sites.
- Configure VLANs in switch net to collect buildings as they are configured today. Then put them on new router as is.
- Do same thing at HSC as UPC in the two collector sites there.
  - Redundant Cisco 7500 to collect networks from two switches.
  - Run HSSI interface to UPC, run backup 10 Mb/s link to UPC.
  - If ‘dark fiber’ is available run several 100 Mb/s FDX links to UPC.
- Upgrade remaining public user facilities with switched 10Mbps network and multiple 100 Mb/s feeds (SAL and part of KOH).

USC Phase 1 Plan (Fall 1997)

- Install GSR and / or ATM switch for connection to Calren-2. Attach to new 7500.
- If we can get SMF(single mode fiber) connection to HSC connect its switches into main switch net. Otherwise use new Cisco 7500 there to interface to the main net over current HSSI interface.
- After migration is done 'in place', start recollecting buildings using VLAN technology and reconfiguring based on IPX and Appletalk net restrictions. Make VLANS for groups, like AIS, based on affinity of users. Do same thing on both campuses. Change configuration of all host to take advantage of 'cut through' switching on VLANs.
- Upgrade network "hot spots". Some buildings are seriously bottlenecked today, for many reasons. Upgrade the hubs in some of these facilities to 10 Mb/s switches with 100 Mb/s feeds to core MMacplus switches. In some buildings replace stacked hub feeds to central switch with multiple feeds. Ongoing throughout the year.
- Install 10x faster backbone ring in UCC machine room to interconnect main servers(numeric facilities and data facilities). Current thought is SCI ring between the Sun 4000s/3000s. Part of discussion on replacement systems for RCF (supercomputer).

USC Phase 2 Plan

- Upgrade building feeds to one or more 100 Mb/s and install 10 Mb/s switches in all buildings.
  - Use 10 Mb/s technology due to Cat 3 wire in most campus buildings.
  - Install 100 Mb/s switched or hub based networks where community funds Cat 5 wire installation and increased hub or switch cost.
- Install 1Gbps ethernet trunk connections between the main collector hubs in Fall 97 when cards are available. Use current 100 Mb/s trunk cards to connect to upgraded buildings as 1 Gb/s cards are available.
- Install ATM switch capability on the collector hubs to allow for ATM connections as needed (QoS or other issues will drive this).
- Start an aggressive program to implement VLANs when switched hubs are installed in buildings. VLANs will allow secure grouping of users that are not in the same building.
- Install VLAN switch hardware in administrative facilities to provide secure high speed connections between AIS users and servers.
- Aggressively work to install fiber based connections to north campus housing and academic facilities.

USC Dialup Pool

- Increase/Modernize Dialup Pool
  - 600 New Modems, 56 kB/s, Replace Standalone Terminal servers, Replace T1 feeds with T3 feed(s), consider building long and short term access pools.
  - 56 kB/s upgrade is free on current modems.
  - Consider charging policies.
  - Change to PPP only access.
  - Connect password/access to ‘ubiquitous account’ system.
  - Start ISDN and ADSL experiments as practical, perhaps fee recovery basis.
  - 1/4 lines by Fall 1997 semester 3/4 later in FY.

Beyond Gigabit Ethernet

At the physical layer, very high-performance low-cost fiber-optic physical layers being developed for use in small multimedia workgroup applications.

SCI (IEEE-1596 1992) and HIPPI-6400 (ANSI X3T11 1997) are possible early implementations with multiple Gb/s sustained throughput to a given node.

Highest performance interconnect based on parallel fiber optics. An example of is the HP/USC-POLO effort.

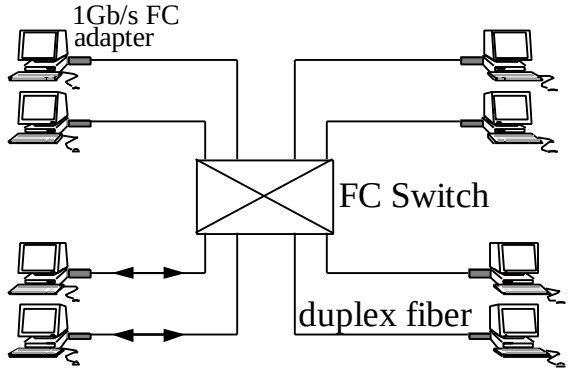
The network data rate and the cost of the network is no longer the bottleneck.

Workgroups supported by networked multiple low-cost end-systems

- Professional campus cluster and workgroup productivity depends on network.
- Investment is in scaleable high-performance network that delivers QoS.
- Multiple low-cost end-systems support individual user needs.
- Multiple low-cost end-systems support telepresence and other group activities.

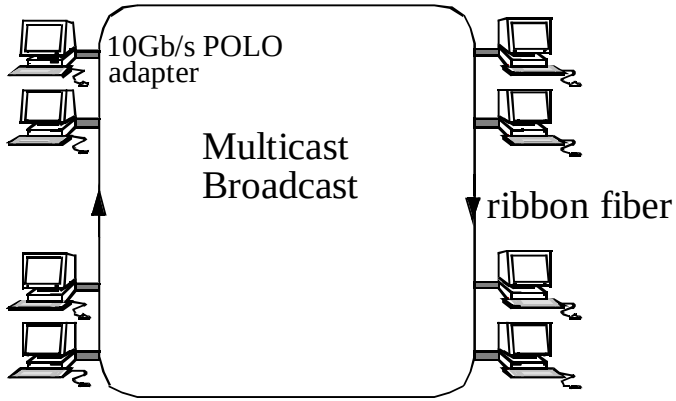
High-Performance Clusters and Workgroups: The POLO Project

0.8 Gb/s at each node
Expensive Fiber Channel (FC) switch
Example: 8 nodes require 16 Tx/Rx modules *and* switch
Difficult multicast/broadcast



POLO network configuration:

More than 10 Gb/s at each node
No expensive switch
Example: 8 nodes only require 8 Tx/Rx modules
POLO adapter cost potentially similar to FC adapter
Natural multicast/broadcast over shared medium

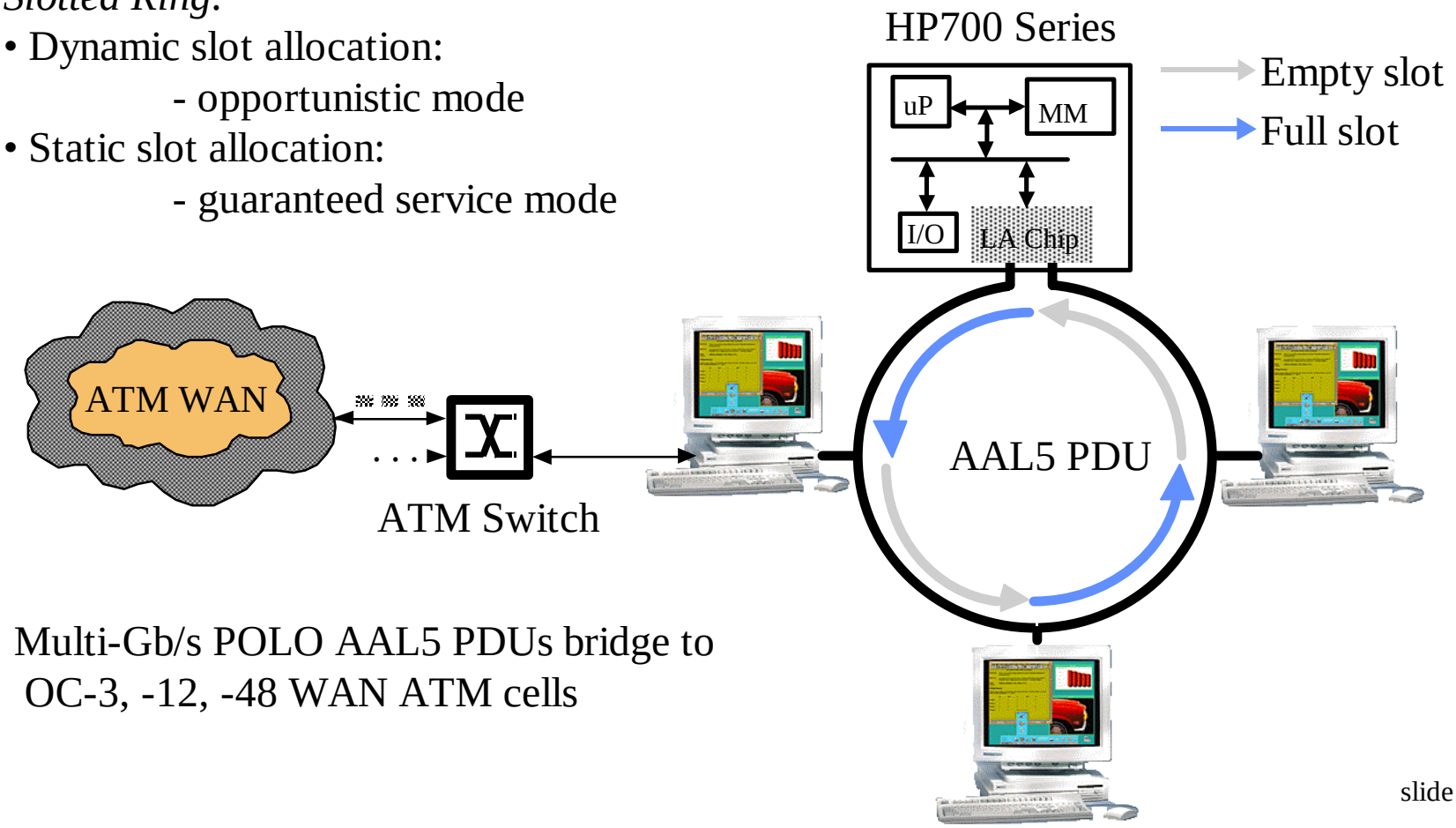


POLO Project

Slotted POLO ring MAC optimized for multimedia applications

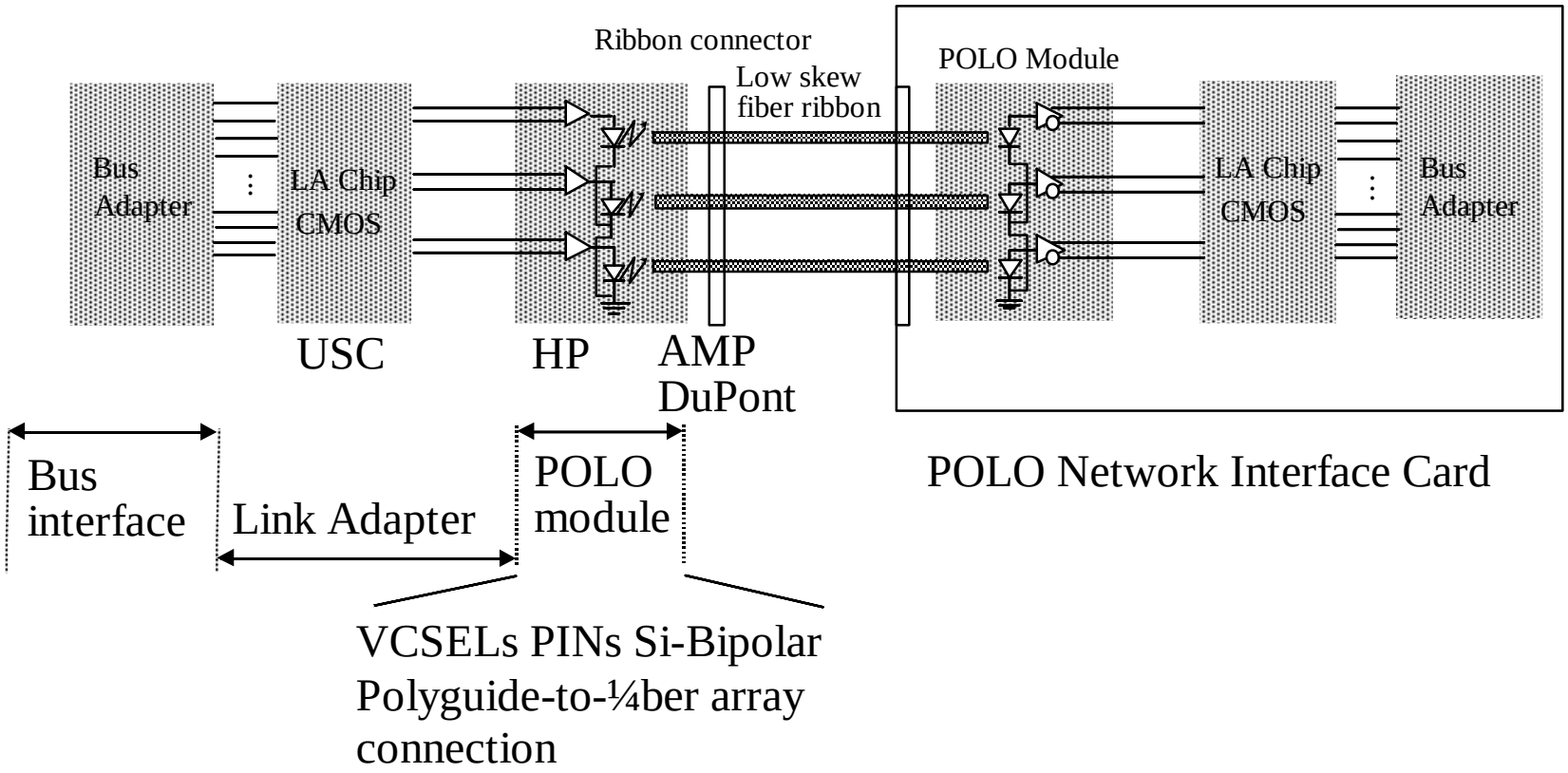
Slotted Ring:

- Dynamic slot allocation:
  - opportunistic mode
- Static slot allocation:
  - guaranteed service mode



Multi-Gb/s POLO AAL5 PDUs bridge to OC-3, -12, -48 WAN ATM cells

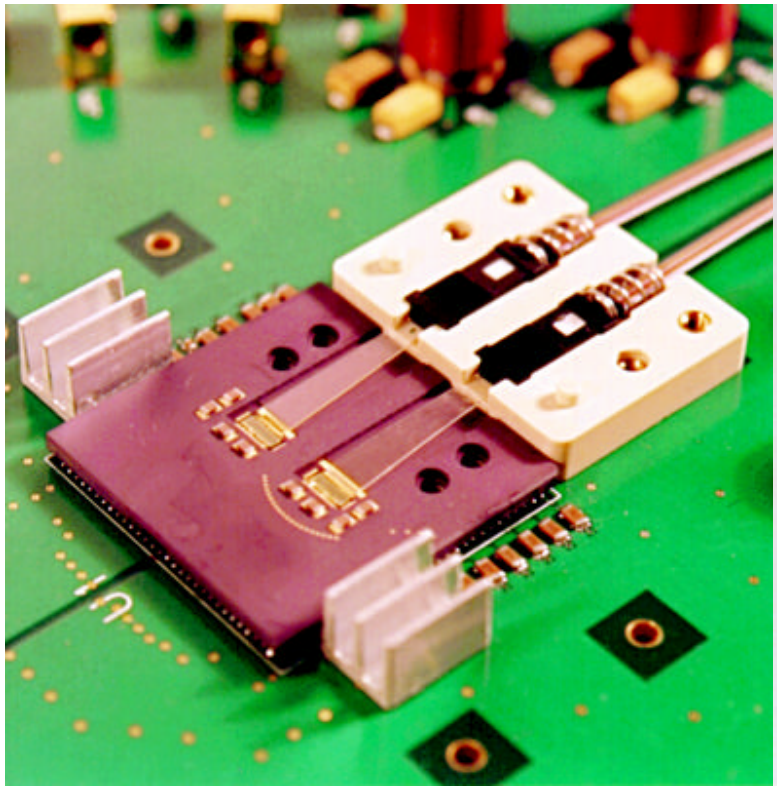
POLO Network Technologies

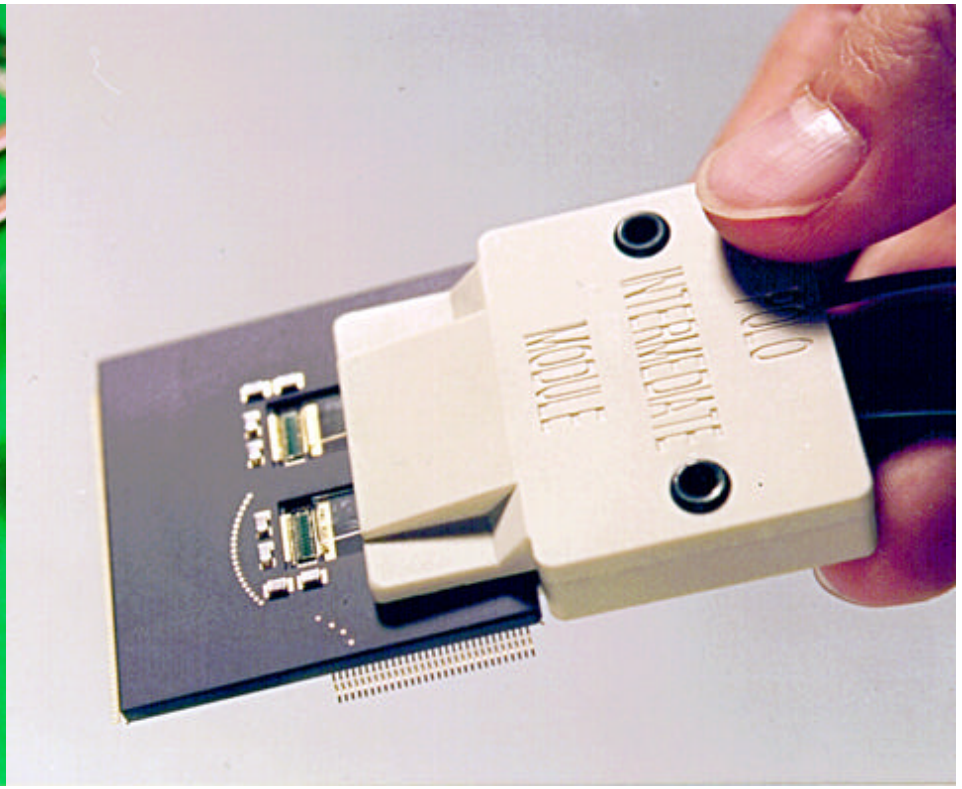


POLO parallel optical interface module fabricated by HP, AMP, DuPont DARPA consortium members. POLO Link Adapter CMOS chip by USC.

HP POLO-1 Tx-Rx Module

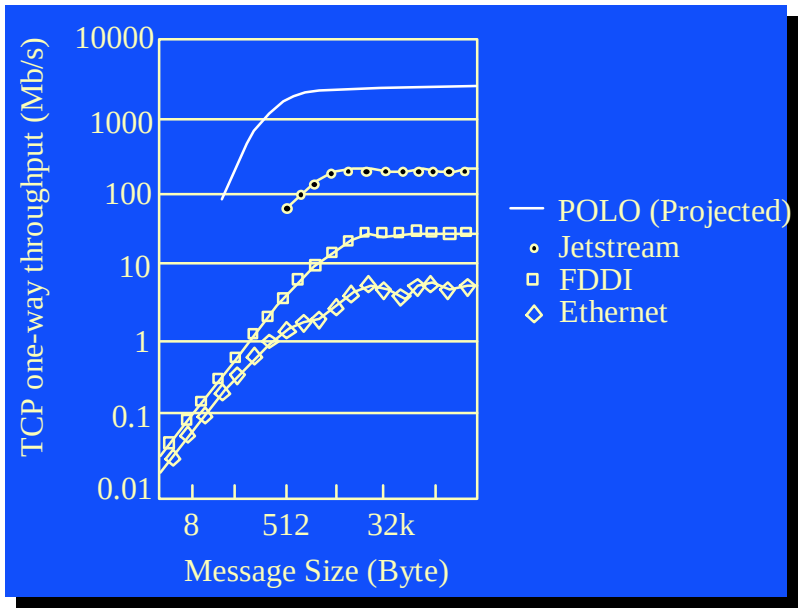
- Uses advanced VCSELs
- 10-wide fiber ribbon
- Plastic wave-guide and fiber array connector components
- Low-cost, multi-Gb/s network physical layer

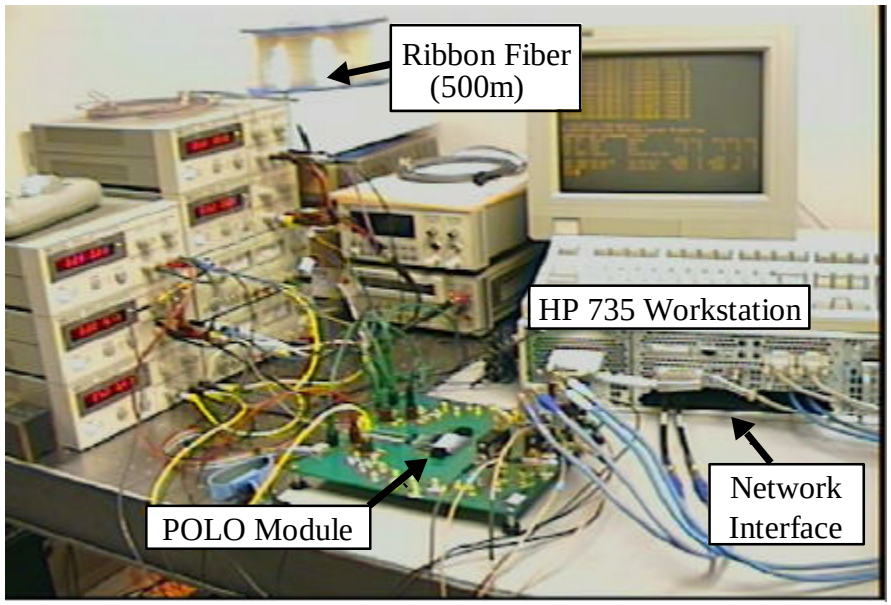




POLO Project

Sustained network throughput and message size





Remote visualization requires high-throughput links

Typical measured one-way TCP Ethernet sustained throughput is 5Mb/s

Typical measured one-way TCP FDDI sustained throughput is 30Mb/s

Jetstream experimental LAN one-copy TCP achieves 200 Mb/s sustained throughput

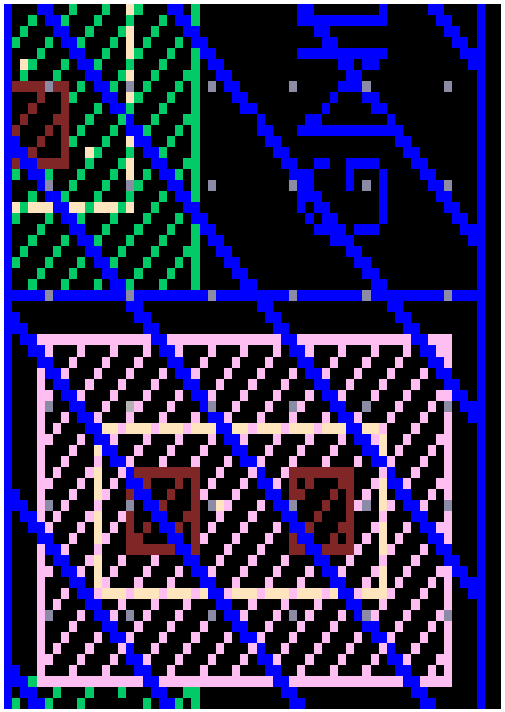
Potential for POLO to achieve greater than 10 times better performance

High Bandwidth Network Applications

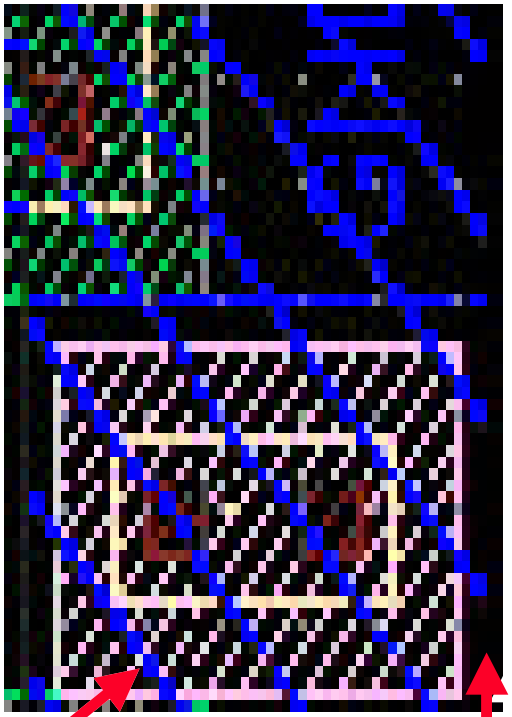
- High bandwidth networks for the Professional Campus
    - TV and movie studios
    - medical campus
- Future applications of high bandwidth networks
    - when and where you will see it
    - impact on multimedia

Networked collaborative CAD applications

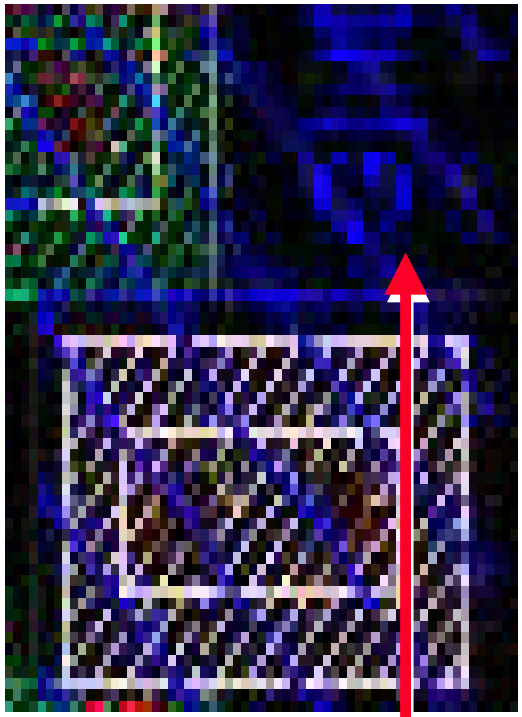
Original image
(uncompressed)



JPEG image
(100% quality)



JPEG image
(50% quality)



Fuzzy line **Missing line** **Unreadable text**

JPEG compression artifacts

JPEG (50%)

Uncompressed

Multiple JPEG compression/decompression editing sessions

Multiple editing using JPEG compression

At random position, cut rectangle out of original uncompressed image

Compress the resulting image using 100% JPEG

Decompress the JPEG image

Replace original rectangle from first step

Artifacts seen around the rectangle

Repeat edit cycle

Cumulative artifacts degrade the image

Why not use lossless compression?

Lossless compression:

- Still requires compression / decompression overhead in hardware / software
- Only a factor 1.4 to 1.6 reduction in image file size

(1:1.6) Majani E. “Lossless compression.” http://www-ias.jpl.nasa.gov/HPCC/eric/node4.html.
(1:1.4) Matrox Inc. “Mathematically Lossless Motion JPEG - Better Than Uncompressed Video.” http://www.matrox.com/videoweb/hot_math.html.
(1:1.5) Oxford University. “JPEG image compression.” Oxford University Libraries Automation Service, http://www.lib.ox.ac.uk/internet/news/faq/archive/jpeg-faq.part1.html.

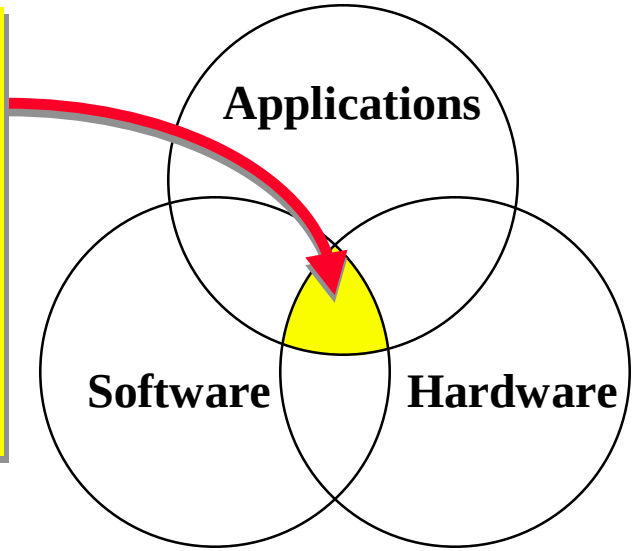
Networked uncompressed data today because:

- Future campus networks have the bandwidth (Gigabit Ethernet, ATM)
- Future PCs / workstations are lower cost and higher performance (memory, disk, CPU, IO bus)

Professional campus network environment

Multimedia applications:
When do networked workstations fail?

- Target the most demanding professional campus applications requiring high-bandwidth lossless image data transport
- Explore weaknesses in system hardware and software and demonstrate solutions



- Professional campus network applications:
  - Studio video editing (television production)
  - Collaborative CAD (engineering)
  - Remote visualization (medical)
  - Graphic arts (advertising)

⇨ The need for artifact-free networked lossless image data

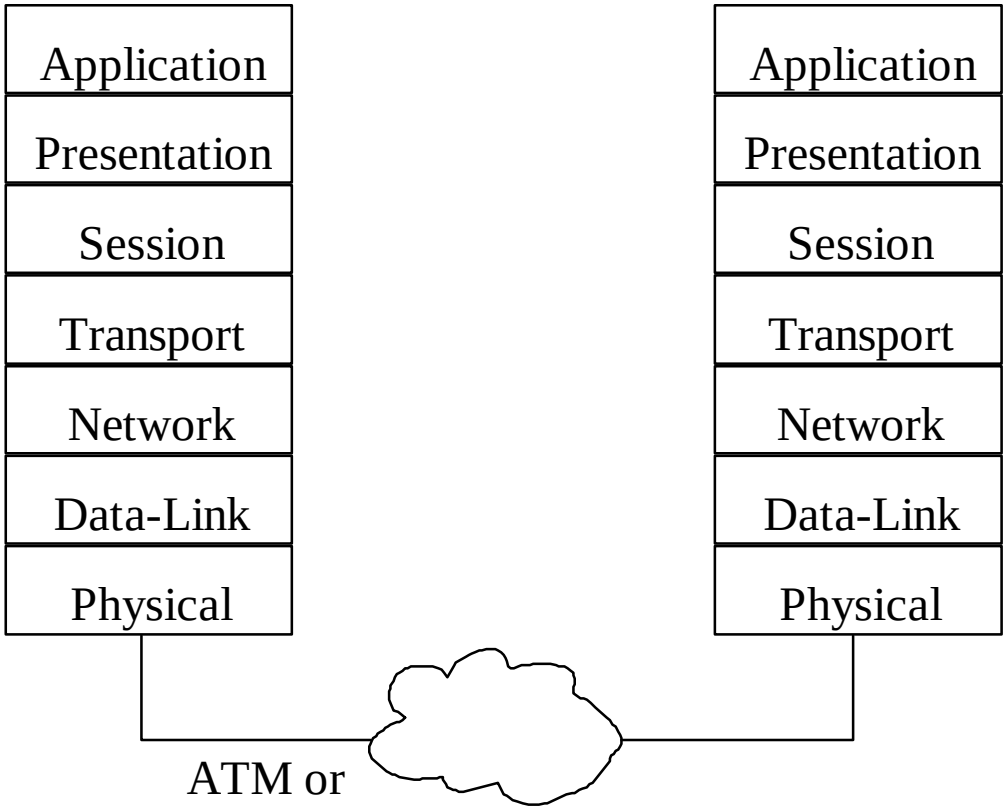
- Enabling technology is integration of:
  - High-performance end-system hardware (fast processors, large memories)
  - Software (NT operating system, device drivers)
  - High-performance network hardware (ATM, Gigabit Ethernet)

High Bandwidth Network Research

- Impact of traffic on performance of high bandwidth networks
- Improved network protocols
- System bottlenecks
    - optimization of hardware and software to maximize Quality-of-Service to end-user

Application-to-application throughput

Initial sustained throughput target
221.18 Mbit/s = 27.65 MB/s
(640 x 480 x 24 x 30 fps)



Growth of CPU - Network performance mismatch

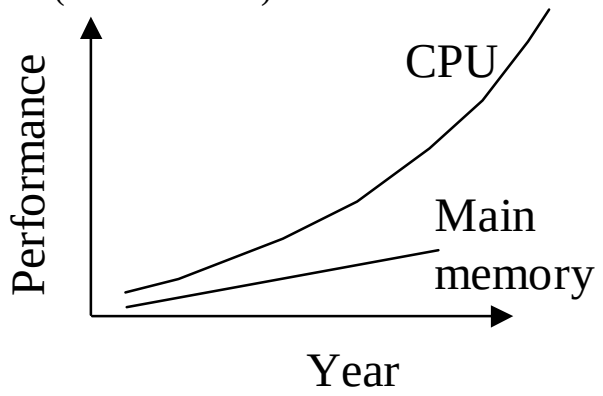
Average CPU clock rate doubles every 18 months

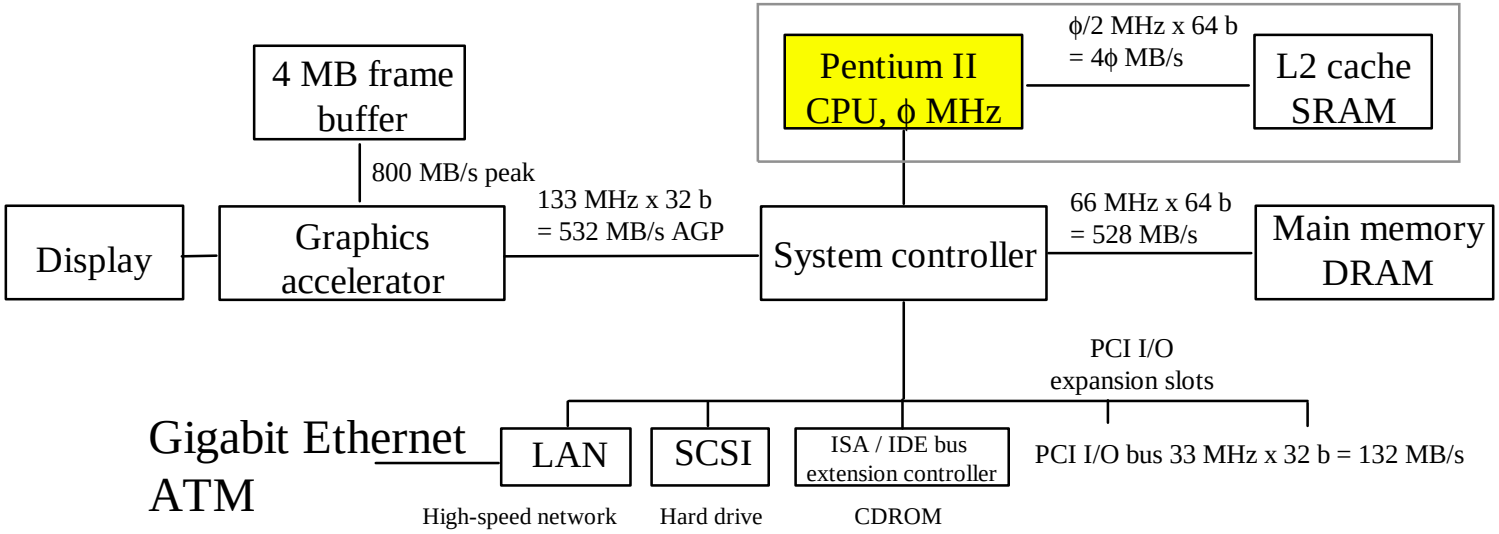
Main memory data transfer speeds increase 10% every 18 months

Network node (end-system) operates at memory to I/O data transfer speeds

High-throughput networked applications forced to minimize number of operations involving main memory

1997 - Jan. Intel Pentium MMX (150 - 233 MHz)
1997 - 2Q AMD K6 MMX (233 - 300 MHz)
1997 - 2Q Intel Pentium II (233 - 300 MHz)
1997 - 3Q Intel Deschutes (300 - 433 MHz)
1998 - Intel Katmai
1999 - Intel Merced



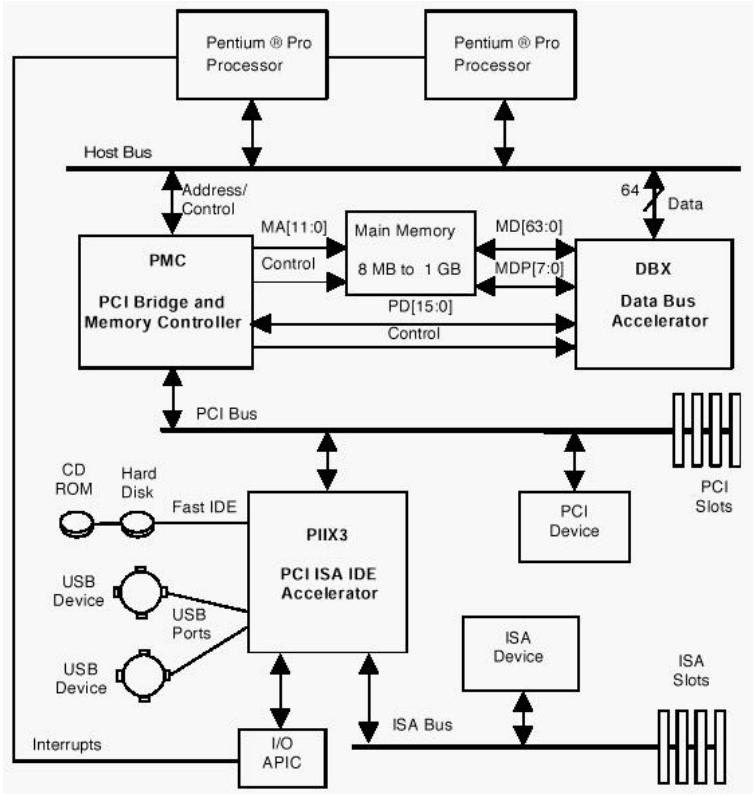


440FX PCI Chipset System Block Diagram

Chipset features include:

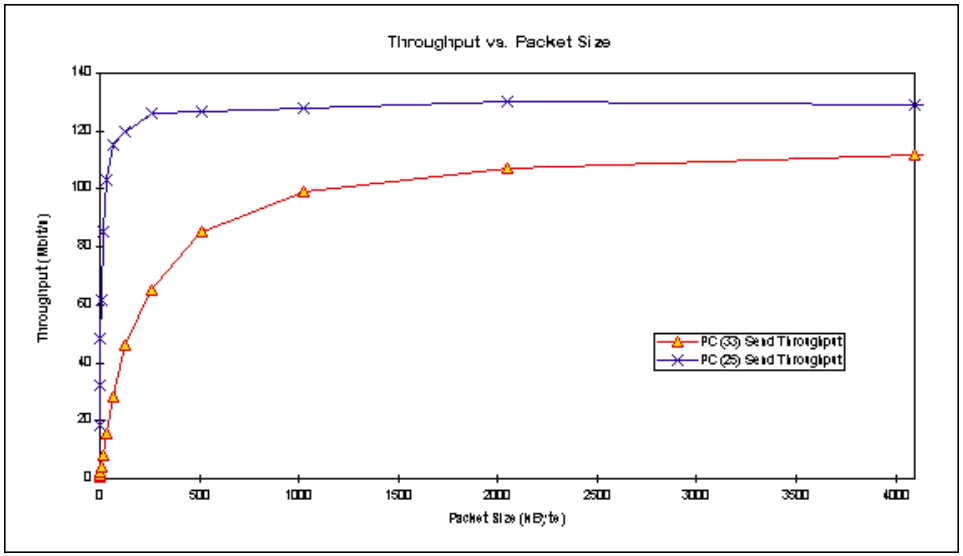
- Support for one or two Pentium Pro Processors at bus frequencies up to 66MHz
- 64/72-bit main memory interface
- 16-bit private bus between DBX and PMC

For more information see:
http://developer.intel.com/design/pcisets/datashts/index.htm



PCI controller bottleneck

Sender throughput using Intel Pentium (100 MHz and 166 MHz) Triton motherboard (50 MHz and 66 MHz) and comparing performance of 82437 FX (25 MHz) and 82437 VX (33 MHz) PCI controller chipset using DMA transfer over AMCC S5933 MatchMaker bridge.
32b x 25 MHz (33 MHz) = 0.80 Gb/s (1.06 Gb/s) maximum possible burst rate.
32 kB packets using machine code (bypassing NT OS) observe 300 Mb/s sustained throughput for VX and 340 Mb/s sustained throughput for FX implementation.
Sustained DMA throughput using Windows NT 4.0 for packets larger than 64 kB is < 140 Mb/s.
Windows NT performs virtual address translation for packets greater than 64 kB. NT device driver improvements include common buffer aligned to 4kB page large enough to accommodate the complete message in a single contiguous region of physical memory.



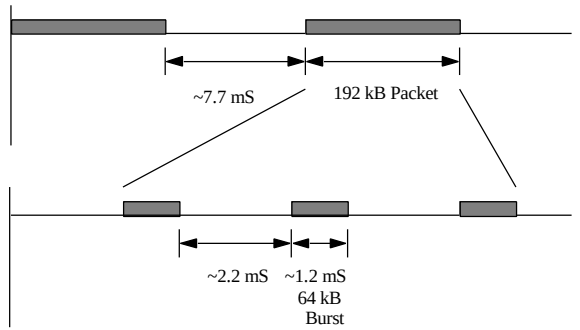


Figure 1: DMA Read for packets >64 kB under Packet DMA in Windows NT The top timeline indicates continuous bursts of 192 kB packets, each of which is broken up into smaller 64 kB chunks.

SGI scaleable cluster and workgroup solution

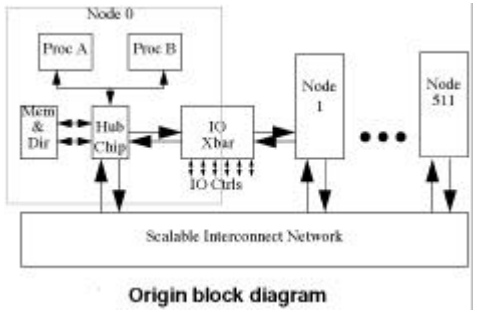
Electrical interconnect:

- 44 signal pins per direction
- up to 5 m electrical cable

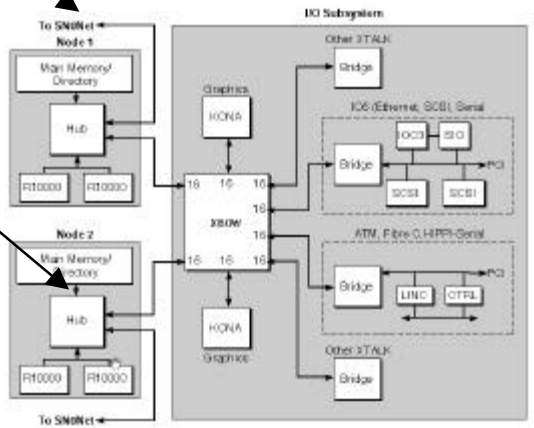
Node-to-node access
0.73 GByte/s peak per direction
0.625 GByte/s sustained per direction

Memory access
0.78 GByte/s peak total
0.78 GByte/s sustained total

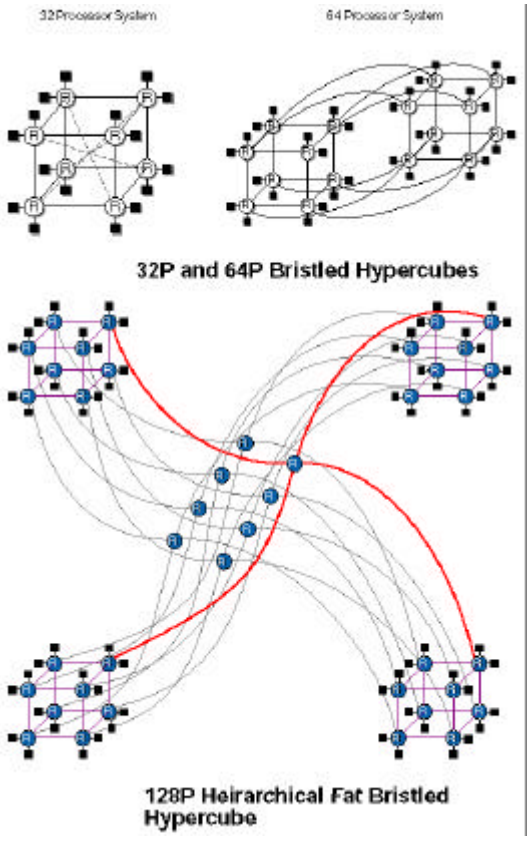
Latency
Pin to pin hub 41 ns
Local memory 310 ns
4P remote memory 540 ns
8P average remote memory 707 ns
16P average remote memory 726 ns
32P average remote memory 773 ns
64P average remote memory 867 ns
128P average remote memory 945 ns



Origin block diagram



Example IO subsystem block diagram



32P and 64P Bristled Hypercubes

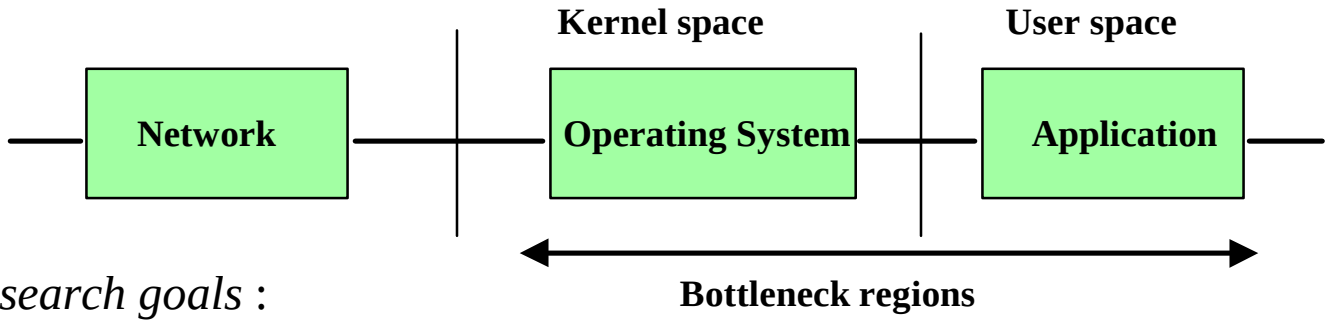
128P Heirarchical Fat Bristled Hypercube

Node-to-node 16 kB page block transfer < 30 µs

High-performance network research issues

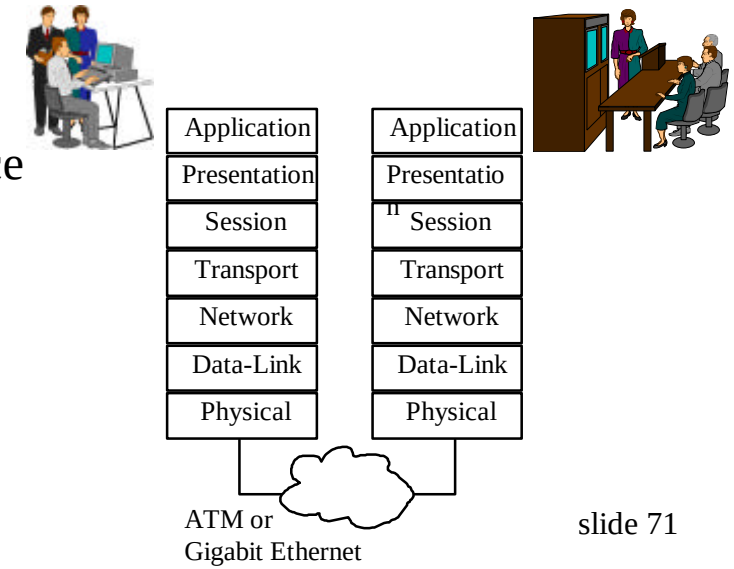
Research problem :

- Mismatch between network and application performance
- End-system bottleneck: hardware, operating system, protocol, application



Research goals :

- End-to-end performance optimization for high-bandwidth applications at network, operating system, and application levels
- Design and implementation of high-performance intelligent API with QoS support for multimedia applications (throughput, jitter, latency)
- Flexible multimedia data manipulation with User-level Multimedia Module (UMM) allowing user control over video data path and video processing



Telepresence in a professional campus environment

High-performance networks: An enabling technology for future interactive professional collaboration

Professional campus network applications:

- Studio video editing
- Collaborative CAD
- Remote visualization
- Graphic art

Remote access to artifact-free networked lossless image data