Lecture 14: Automated Data Libraries & Networks & Interconnect—Introduction

Professor David A. Patterson

Computer Science 252

Fall 1996

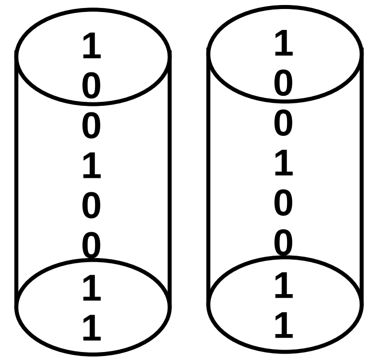
Review: RAID Techniques

- ***Disk Mirroring, Shadowing***

  Each disk is fully duplicated onto its "shadow"

  Logical write = two physical writes

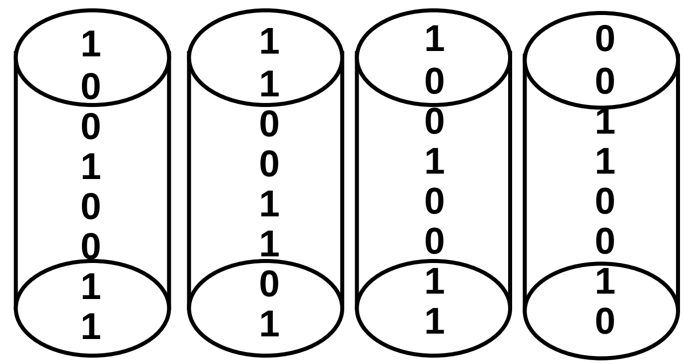
  100% capacity overhead



- ***Parity Data Bandwidth Array***

  Parity computed horizontally

  Logically a single high data bw disk



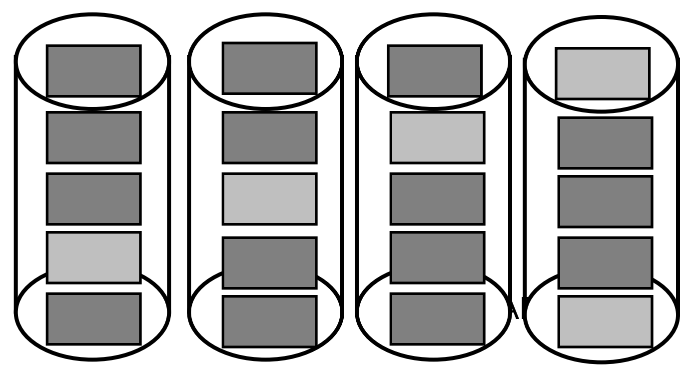
- ***High I/O Rate Parity Array***

  Interleaved parity blocks

  Independent reads and writes

  Logical write = 2 reads + 2 writes

  Parity + Reed-Solomon codes



Review: RAID

RAID sales, “The Independent RAID Report” May/June 1994, p. 15 (dwilmot@crl.com, 510-938-7425)

- **1993: $3.4 billion on 214,667 arrays (≈ $15,000 / RAID)**
- **1996 forecast: $11 billion**
- **1997 forecast: $13 billion on 837,155 units**
  - **Source: DISK/TREND, 5/94 (415-961-6209)**

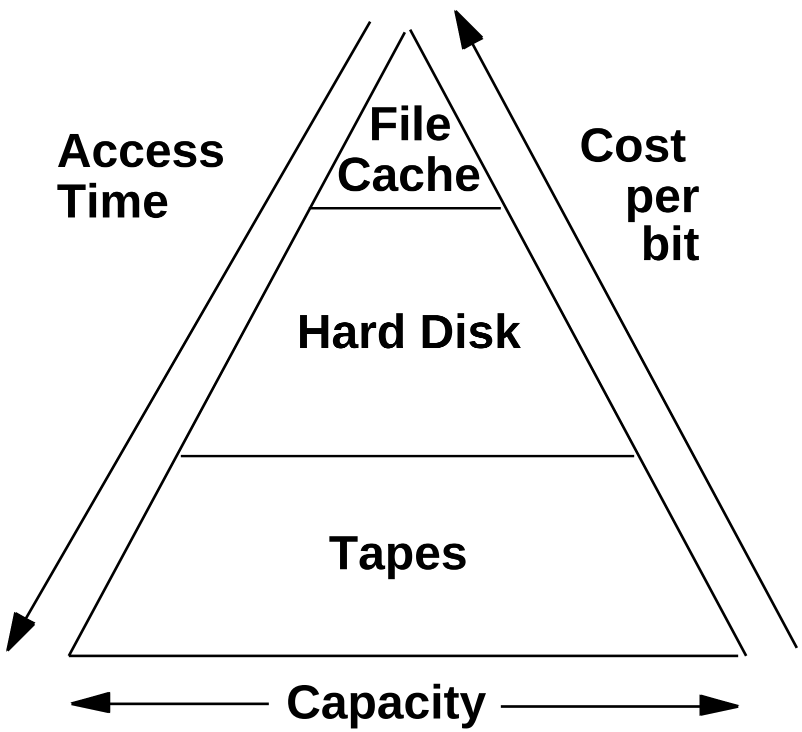
Summary: I/O Benchmarks

- **Scaling to track technological change**
- **TPC: price performance as nomalizing configuration feature**
- **Auditing to ensure no foul play**
- **Througput with restricted response time is normal measure**

Review: Storage System Issues

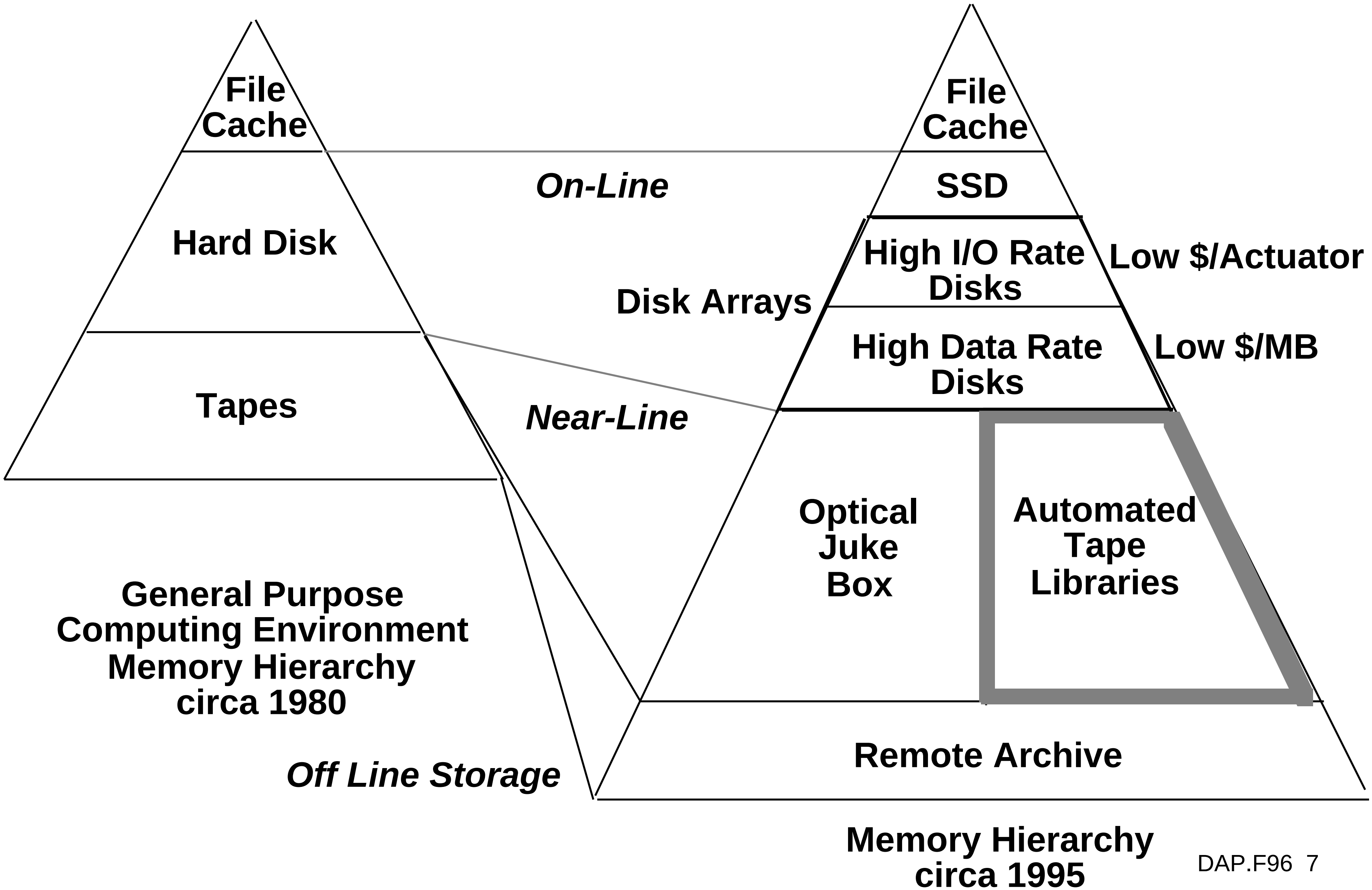
- Historical Context of Storage I/O
- Storage I/O Performance Measures
- Secondary and Tertiary Storage Devices
- A Little Queuing Theory
- Processor Interface Issues
- I/O & Memory Buses
- RAID
- ABCs of UNIX File Systems
- I/O Benchmarks
- Comparing UNIX File System Performance
- Tertiary Storage Possbilities

Memory Hierarchies



General Purpose
Computing Environment
Memory Hierarchy
circa 1980

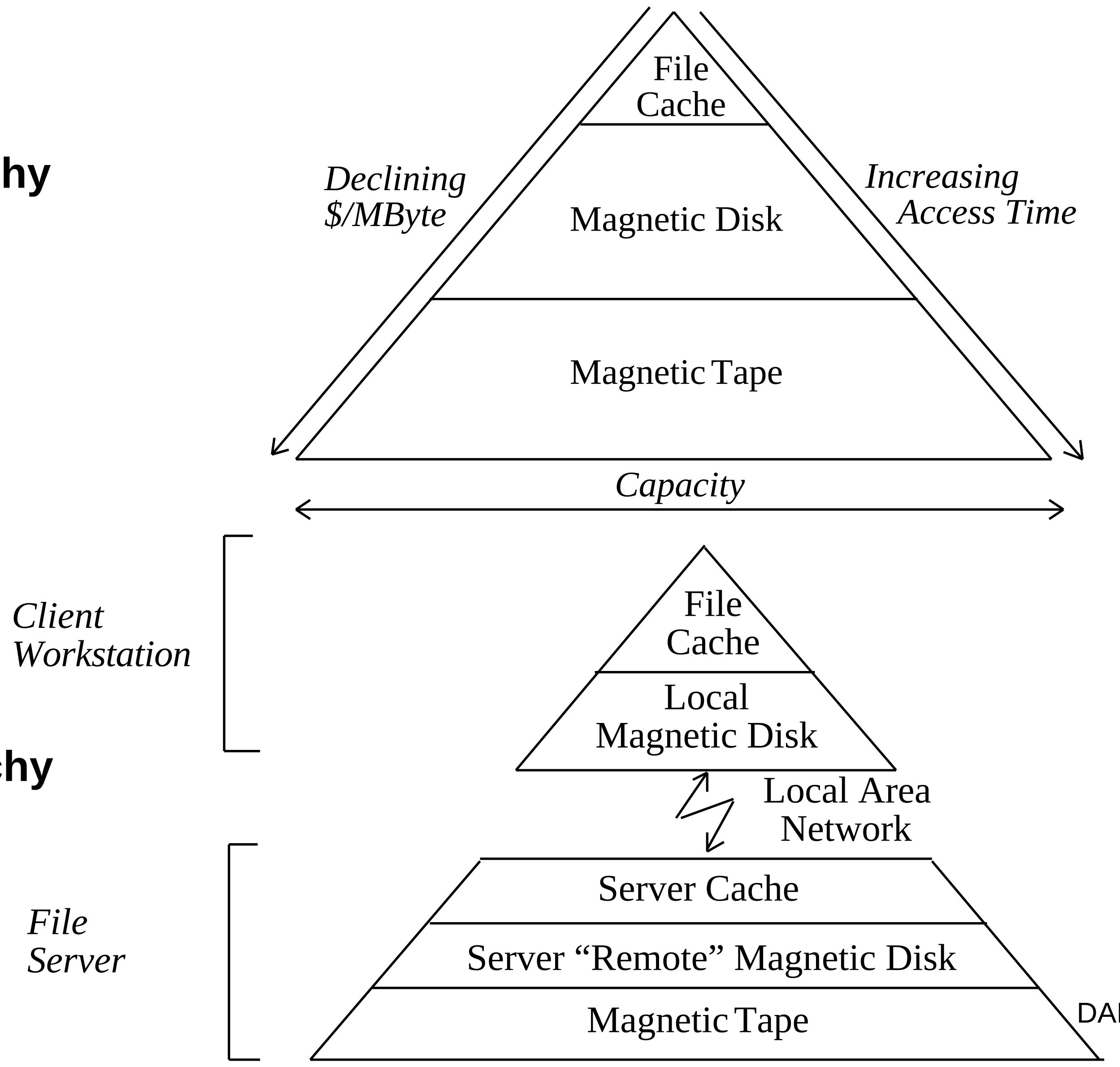
Memory Hierarchies



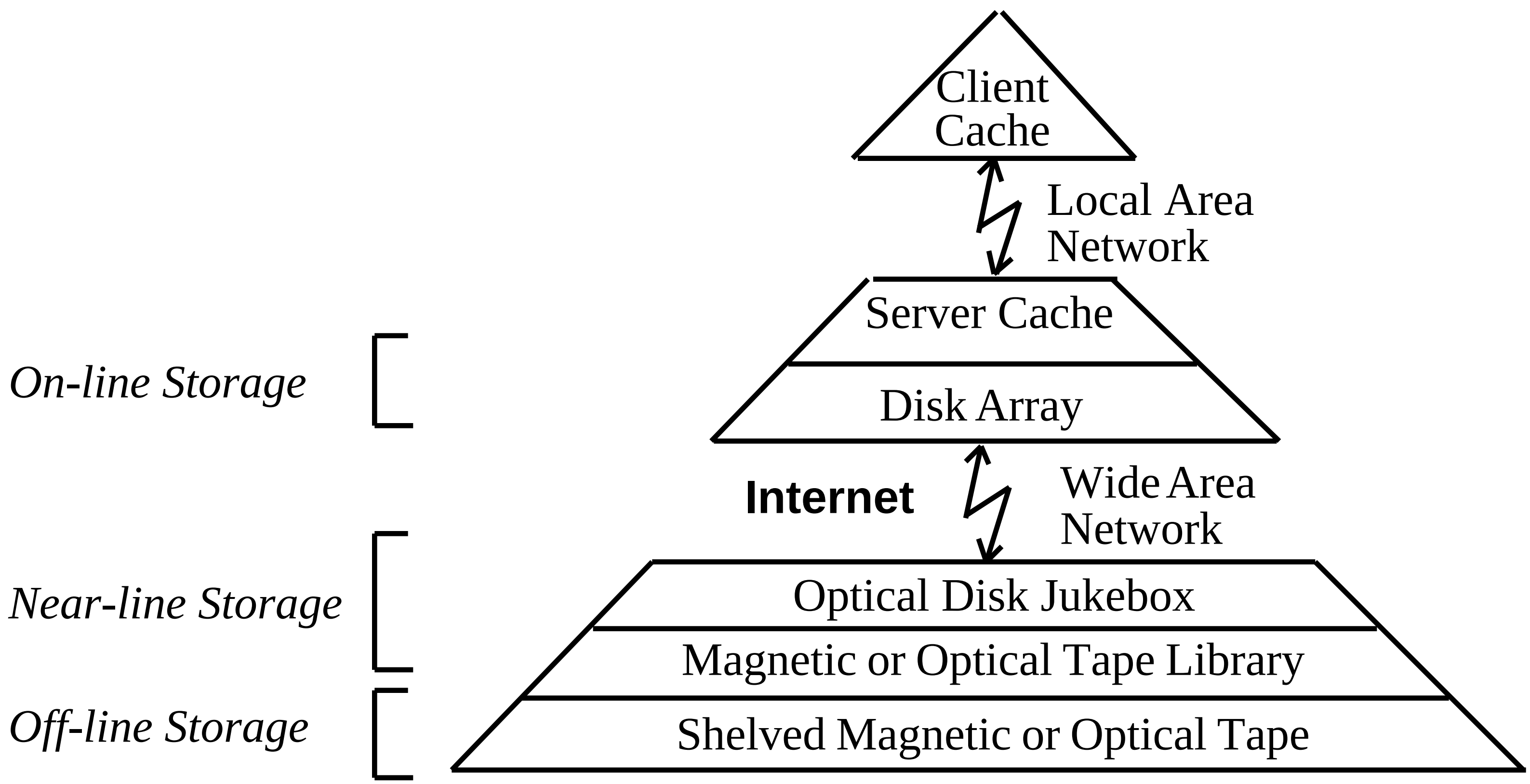
Storage Trends: Distributed Storage

Storage Hierarchy circa 1980

Storage Hierarchy circa 1990



Storage Trends: Wide-Area Storage



Typical Storage Hierarchy, circa 1995

Conventional disks replaced by disk arrays

Near-line storage emerges between disk and tape

What's All This About Tape?

Tape is used for:

- **Backup Storage for Hard Disk Data**

  Written once, very infrequently (hopefully never!) read

- **Software Distribution**

  Written once, read once

- **Data Interchange**

  Written once, read once

- **File Retrieval**

  Written/Rewritten, files occasionally read

  Near Line Archive

  Electronic Image Management

Relatively New Application For Tape

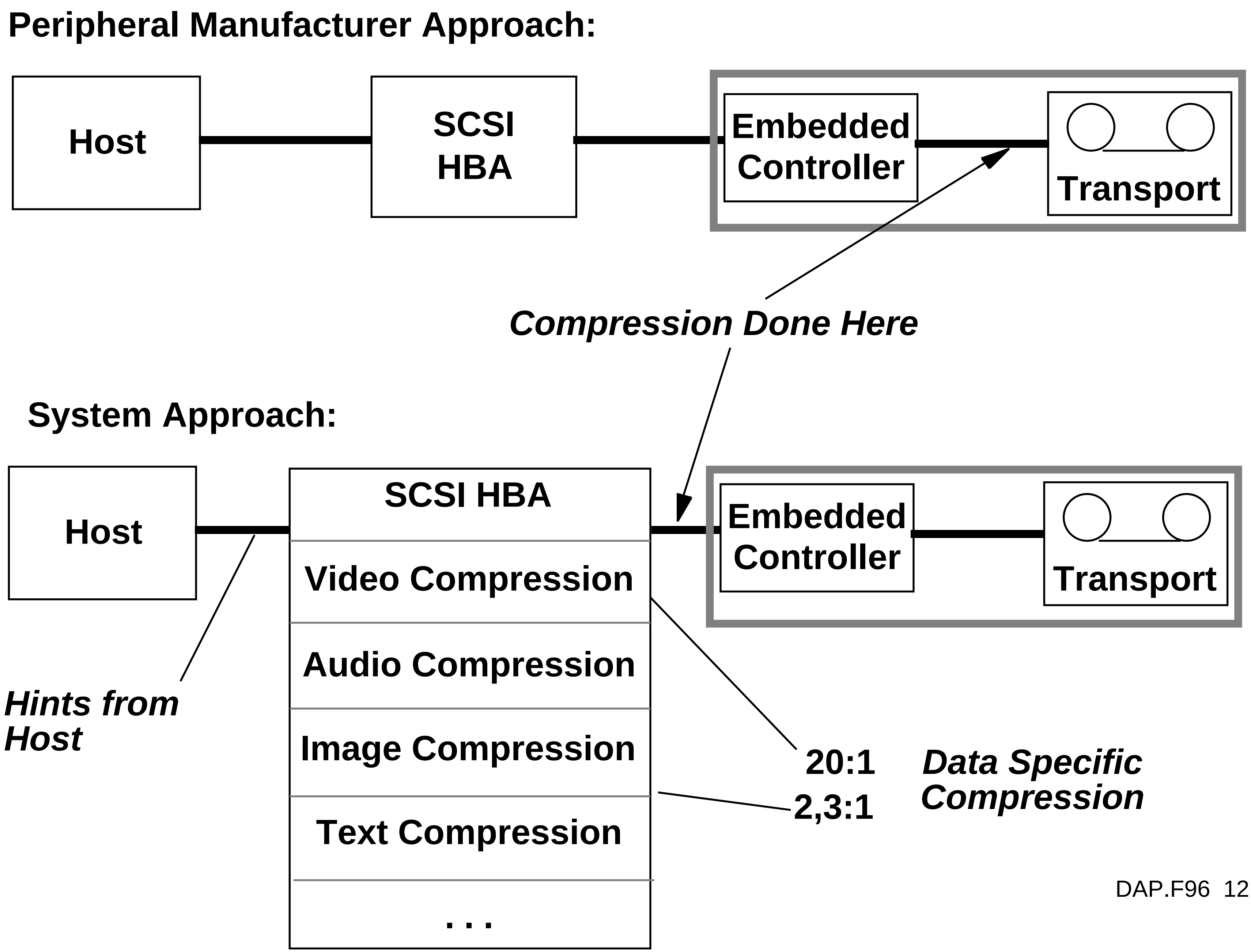
Alternative Data Storage Technologies

Technology	Cap (MB)	BPI	TPI	BPI*TPI (Million)	Data Xfer (KByte/s)	Access Time
Conventional Tape:						
Reel-to-Reel (.5")	140	6250	18	0.11	549	minutes
Cartridge (.25")	150	12000	104	1.25	92	minutes
Helical Scan Tape:						
VHS (.5")	2500	17435	650	11.33	120	minutes
Video (8mm)*	2300	43200	819	35.28	246	minutes
DAT (4mm)**	1300	61000	1870	114.07	183	20 seconds
Disk:						
Hard Disk (5.25")	760	30552	1667	50.94	1373	20 ms
Floppy Disk (3.5")	2	17434	135	2.35	92	1 second
CD ROM (3.5")	540	27600	15875	438.15	183	1 second

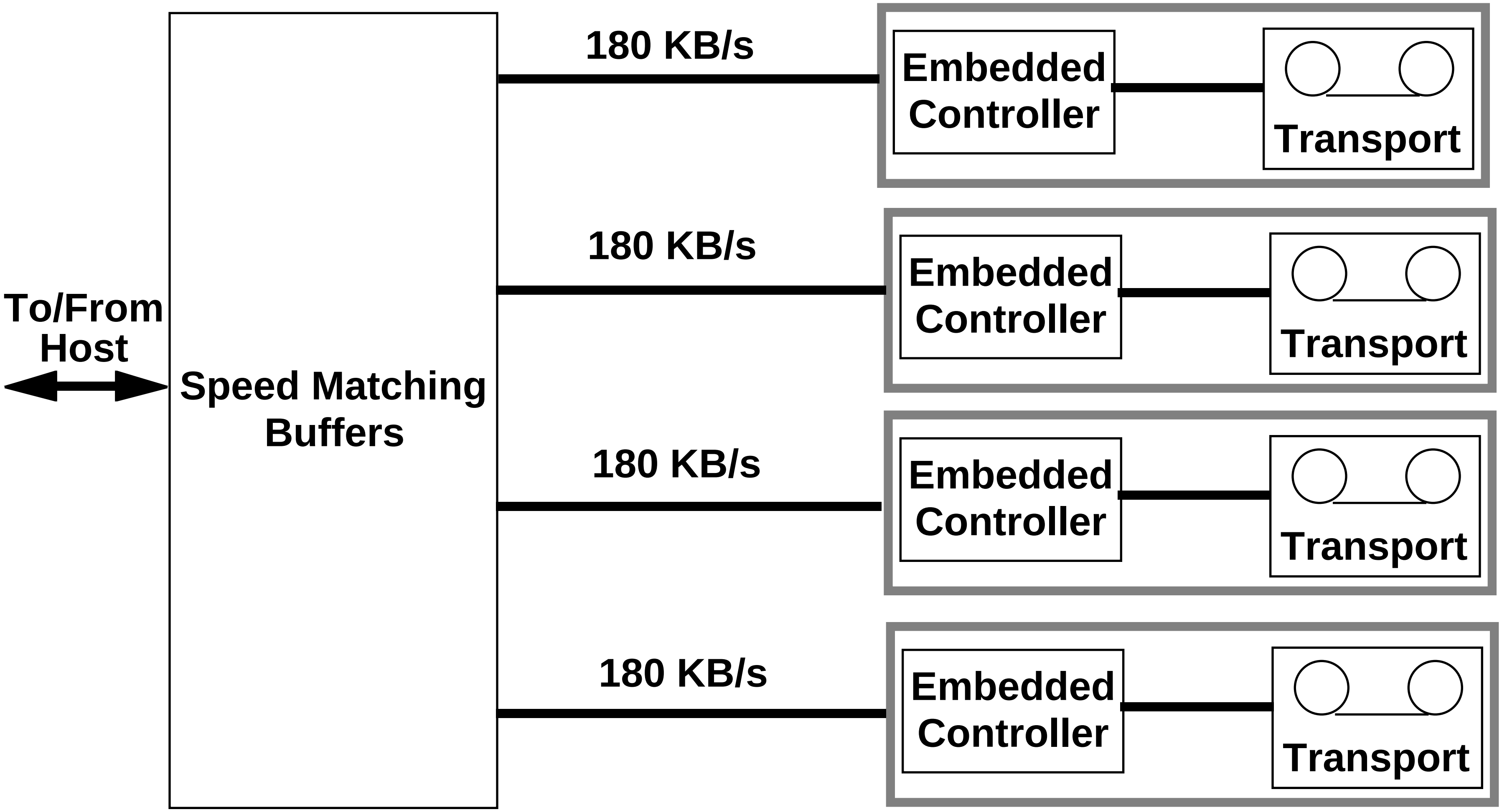
* Second Generation 8mm: 5000 MB, 500KB/s

** Second Generation 4mm: 10000 GB

Data Compression Issues



Striped Tape

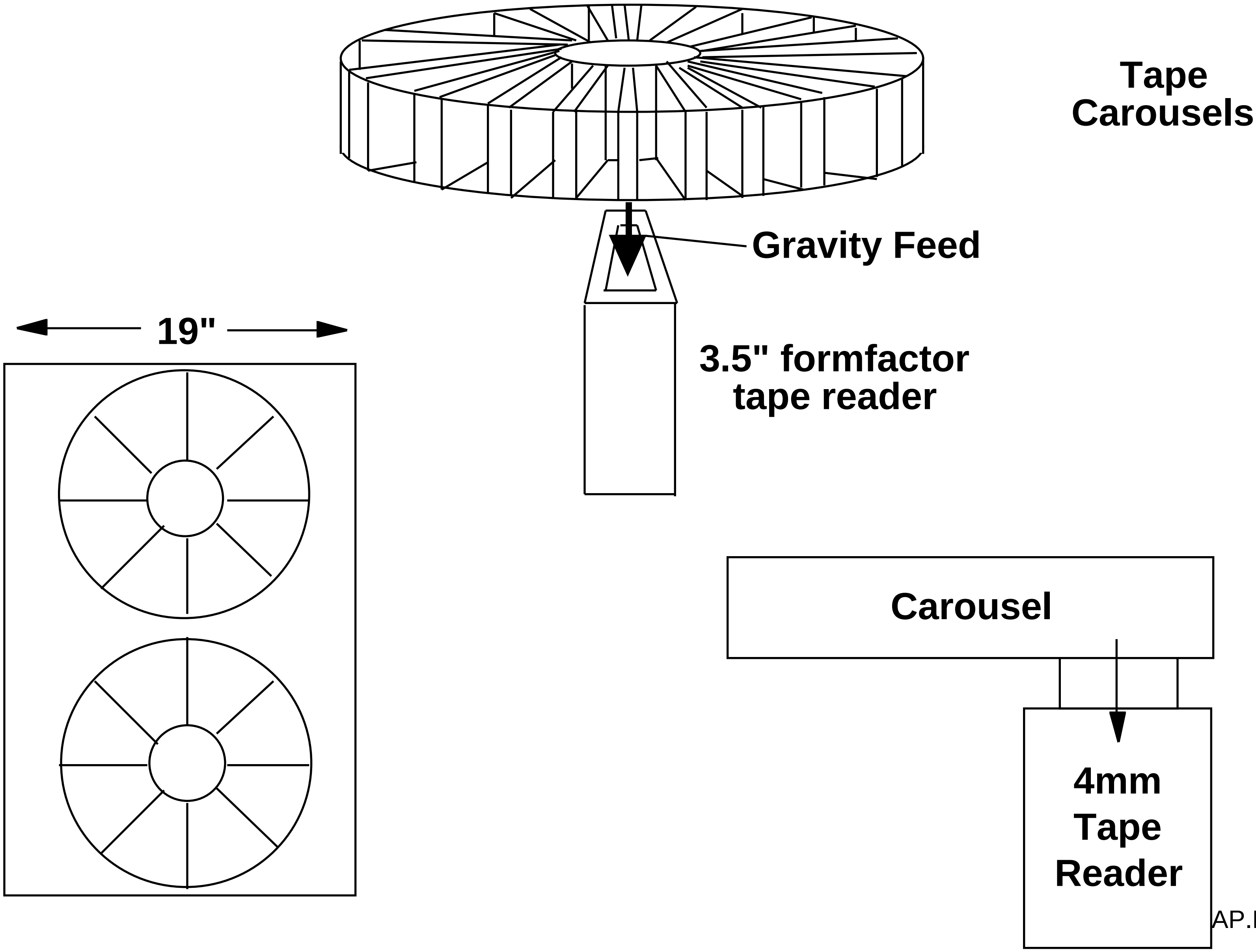


Challenges:

- Difficult to logically synchronize tape drives
- Unpredictable write times

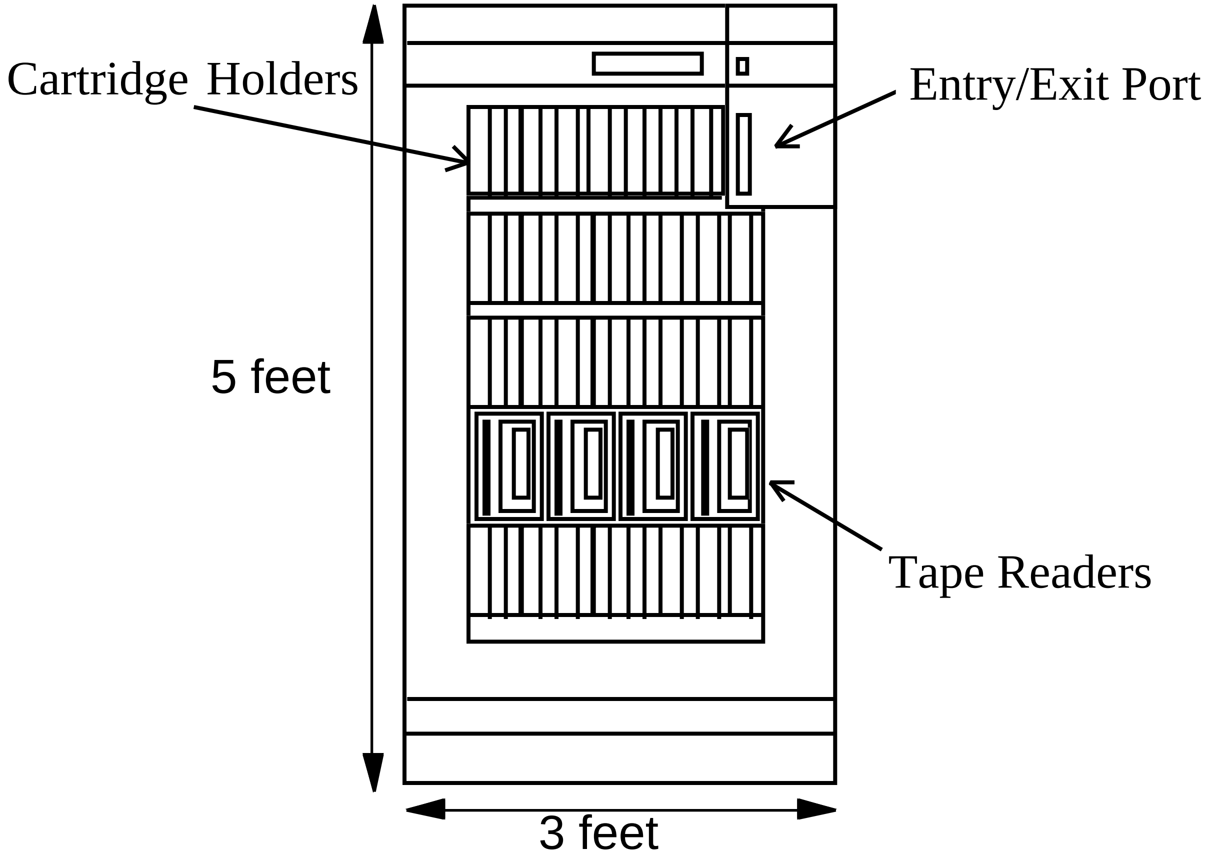
  R after W verify, Error Correction Schemes, N Group Writing, Etc.

Automated Media Handling



MSS: Automated Tape Library

EXB-120



- **116 x 5 GB 8 mm tapes = 0.6 TBytes (1991)**
- **4 tape readers 1991, 8 half height readers now**
- **4 x .5 MByte/second = 2 MBytes/s**
- **$40,000 O.E.M. Price**
- **1995: 3 TBytes; 2000: 9 TBytes**

CS 252 Administrivia

- Homework on Chapter 6 due Monday10/21 at 5PM in 252 box, done in pairs:
  - Exercises 6.5, 6.16, 6.17
- Project Survey #2 due Wednesday 10/23 in class

Open MSS Research Issues

- Hardware/Software attack on very large storage systems
  - File system extensions to handle terabyte sized file systems
  - Storage controllers able to meet bandwidth and capacity demands
- Compression/decompression between secondary and tertiary storage
  - Hardware assist for on-the-fly compression
  - Application hints for data specific compression
  - More effective compression over large buffered data
  - DB indices over compressed data
- Striped tape: is large buffer enough?
- Applications: Where are the Terabytes going to come from?
  - Image Storage Systems
  - Personal Communications Network multimedia file server

MSS: Applications of Technology
Robo-Line Library

Books/Bancroft x Pages/book x bytes/page = Bancroft

Books/Bancroft	Pages/book	bytes/page	Bancroft
372,910	400	4000	= 0.54 TB

Full text Bancroft Near Line = 0.5 TB;

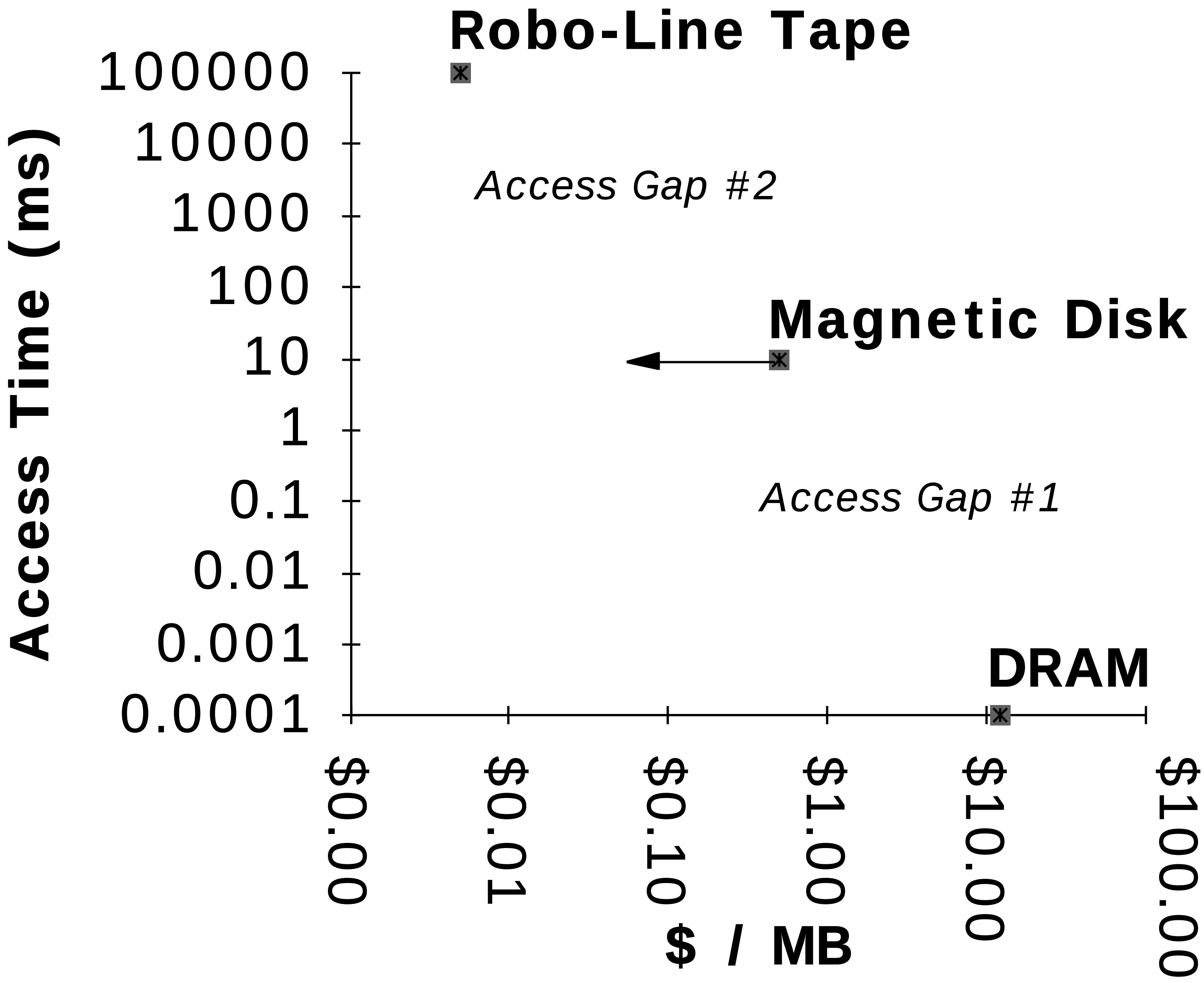
Pages images ≈ 20 TB

Predict: "RLB" (Robo-Line Bancroft) = $250,000

Bancroft costs:

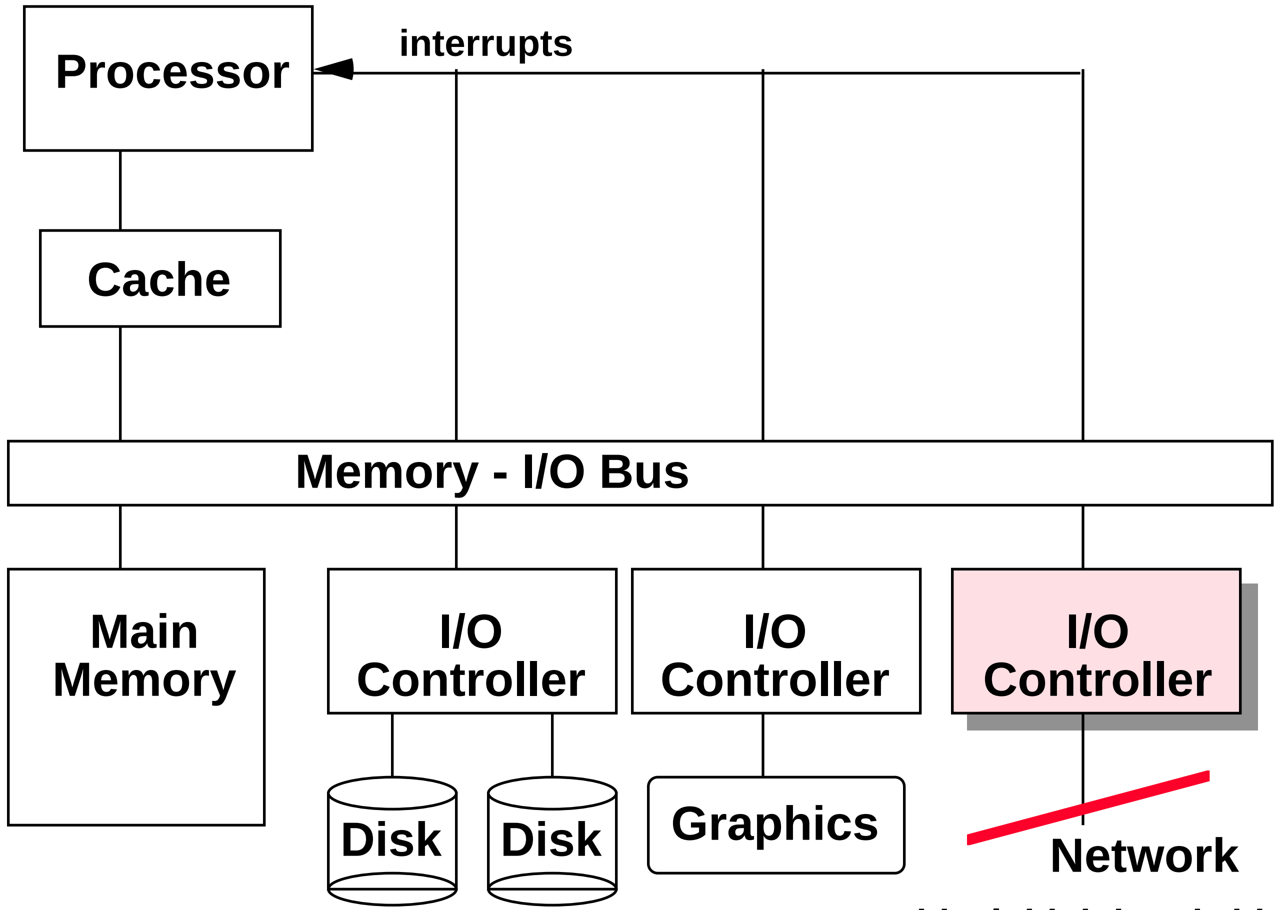
Catalogue a book:	$20 / book
Reshelve a book:	$1/ book
% new books purchased per year never checked out:	20%

MSS: Summary



Robo-Line Tape
100000
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
Access Time (ms)
Access Gap #2
Magnetic Disk
Access Gap #1
DRAM
$0.00
$0.01
$0.10
$1.00
$10.00
$100.00
$ / MB

I/O to External Devices and Other Computers

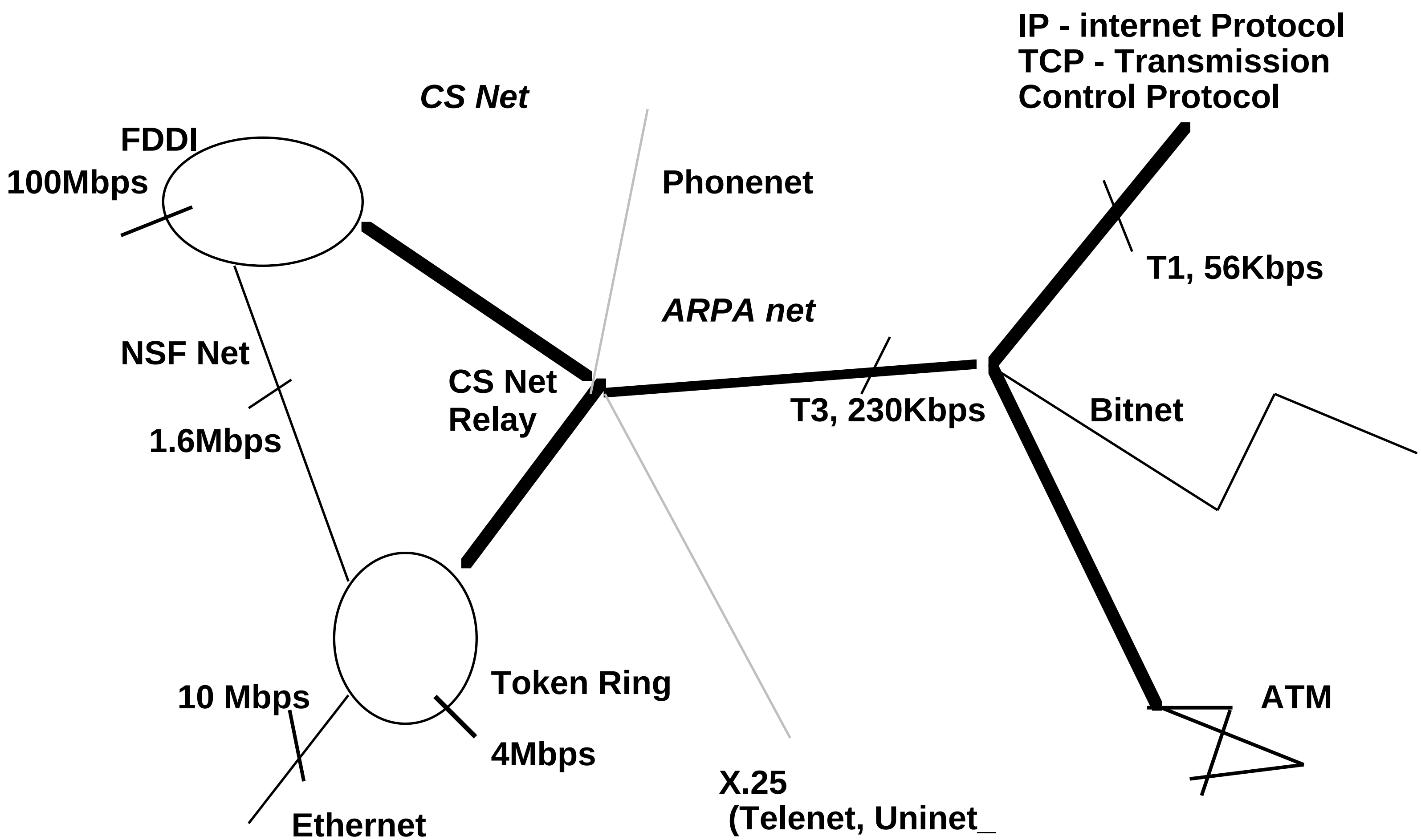


ideal: high bandwidth, low latency

Networks

- **Goal: Communication between computers**
- **Eventual Goal: treat collection of computers as if one big computer, distributed resource sharing**
- **Theme: Different computers must agree on many things**
  - **Overriding importance of standards and protocols**
- **Warning: Terminology-rich environment**

Example Major Networks



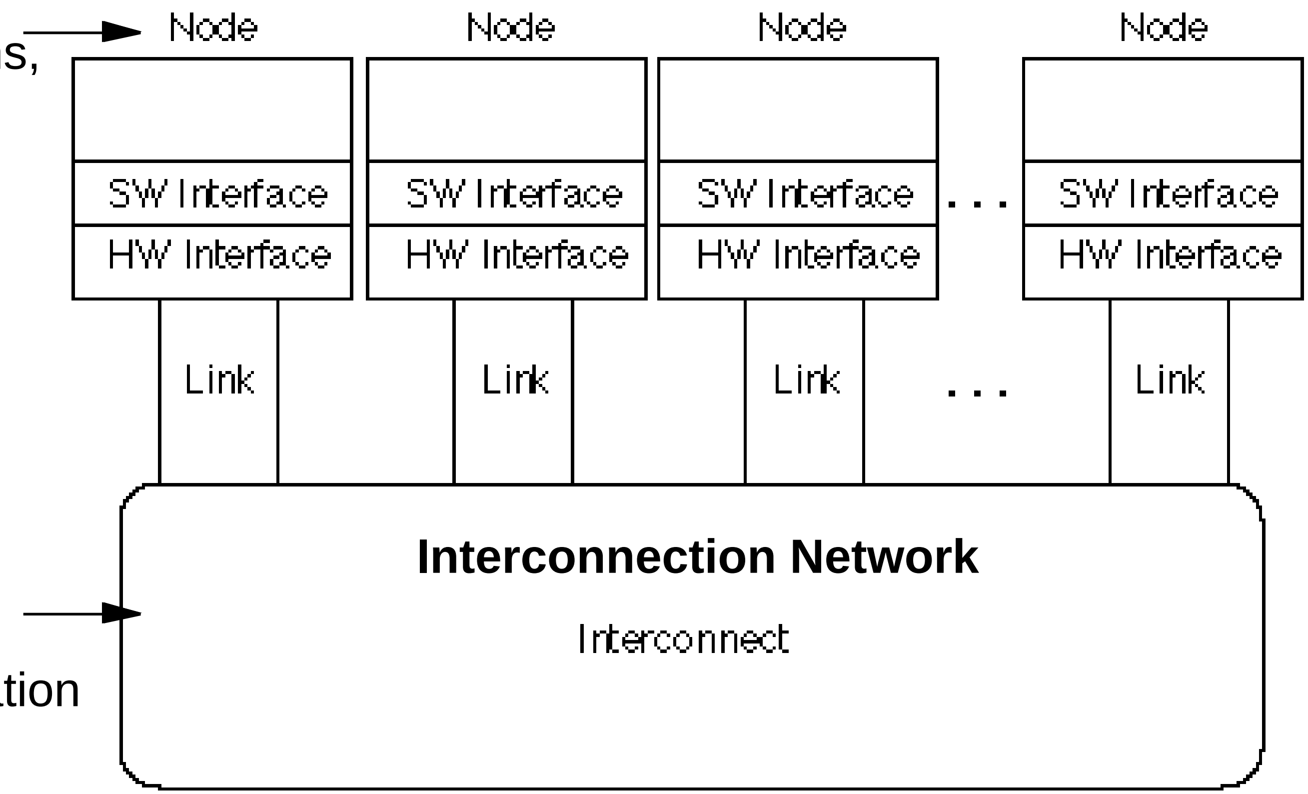
Networks

- **Facets people talk a lot about:**
  - direct (point-to-point) vs. indirect (multi-hop)
  - topology (e.g., bus, ring, DAG)
  - routing algorithms
  - switching (aka multiplexing)
  - wiring (e.g., choice of media, copper, coax, fiber)
- **What really matters:**
  - latency
  - bandwidth
  - cost
  - reliability

Interconnections (Networks)

- **Examples:**
  - **MPP networks (SP2): 1000s nodes; ≤ 25 meters per link**
  - **Local Area Networks (Ethernet): 100s nodes; ≤ 1000 meters**
  - **Wide Area Network (ATM): 1000s nodes; ≤ 5,000,000 meters**

a.k.a.
end systems,
hosts



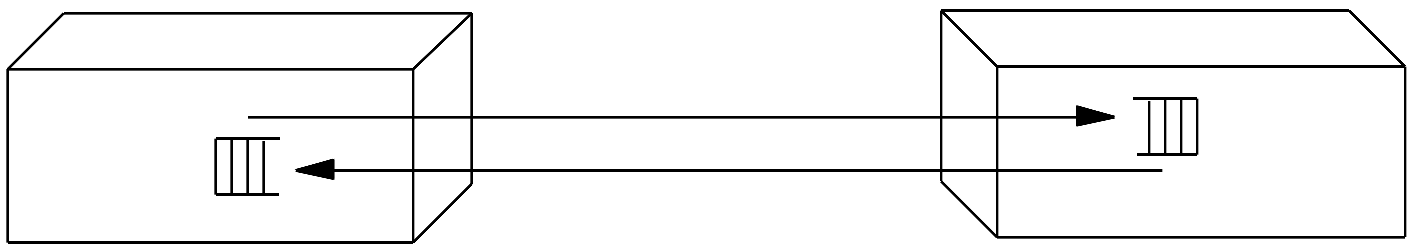
a.k.a.
network,
communication
subnet

More Network Background

- Connection of 2 or more networks: **Internetworking**
- 3 cultures for 3 classes of networks
  - MPP: performance, latency and bandwidth
  - LAN: workstations, cost
  - WAN: telecommunications, phone call revenue
- Try for single terminology
- Motivate the complexity incrementally

ABCs of Networks

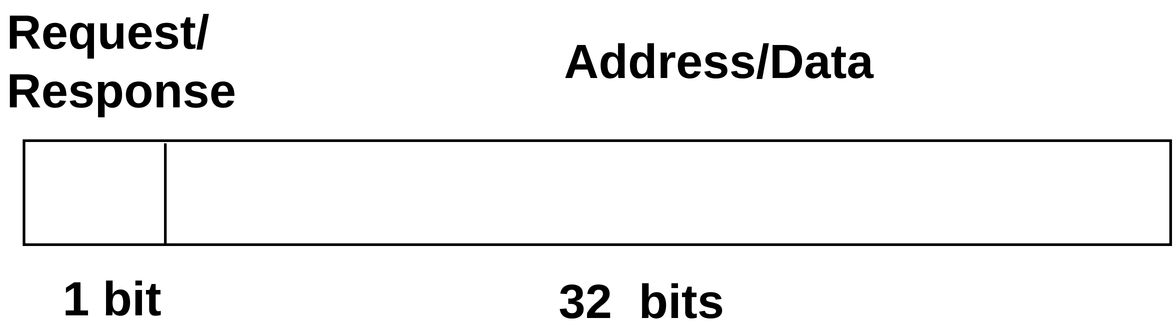
- **Starting Point: Send bits between 2 computers**



- **Queue (FIFO) on each end**
- **Information sent called a message**
- **Can send both ways ("Full Duplex")**
- **Rules for communication? "protocol"**
  - **Inside a computer:**
    - » **Loads/Stores: Request (Address) & Response (Data)**
    - » **Need Request & Response signaling**

A Simple Example

- **What is the format of mesage?**
  - **Fixed? Number bytes?**



0: Please send data from Address
1: Packet contains data corresponding to request

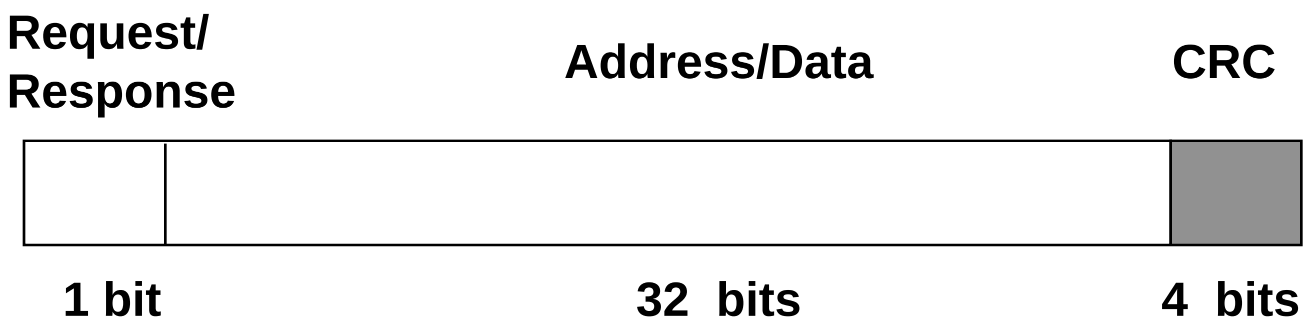
- **Header/Trailer: information to deliver a message**
- **Payload: data in message (1 word above)**

Questions About Simple Example

- **What if more than 2 computers want to communicate?**
  - **Need computer address field (destination) in packet**
- **What if packet is garbled in transit?**
  - **Add error detection field in packet (e.g., CRC)**
- **What if packet is lost?**
  - **More elaborate protocols to detect loss (e.g., NAK, ARQ, time outs)**
- **What if multiple processes/machine?**
  - **Queue per process to provide protection**
- **Questions such as these lead to more complex protocols and packet formats**

A Simple Example Revisted

- **What is the format of packet?**
  - **Fixed? Number bytes?**

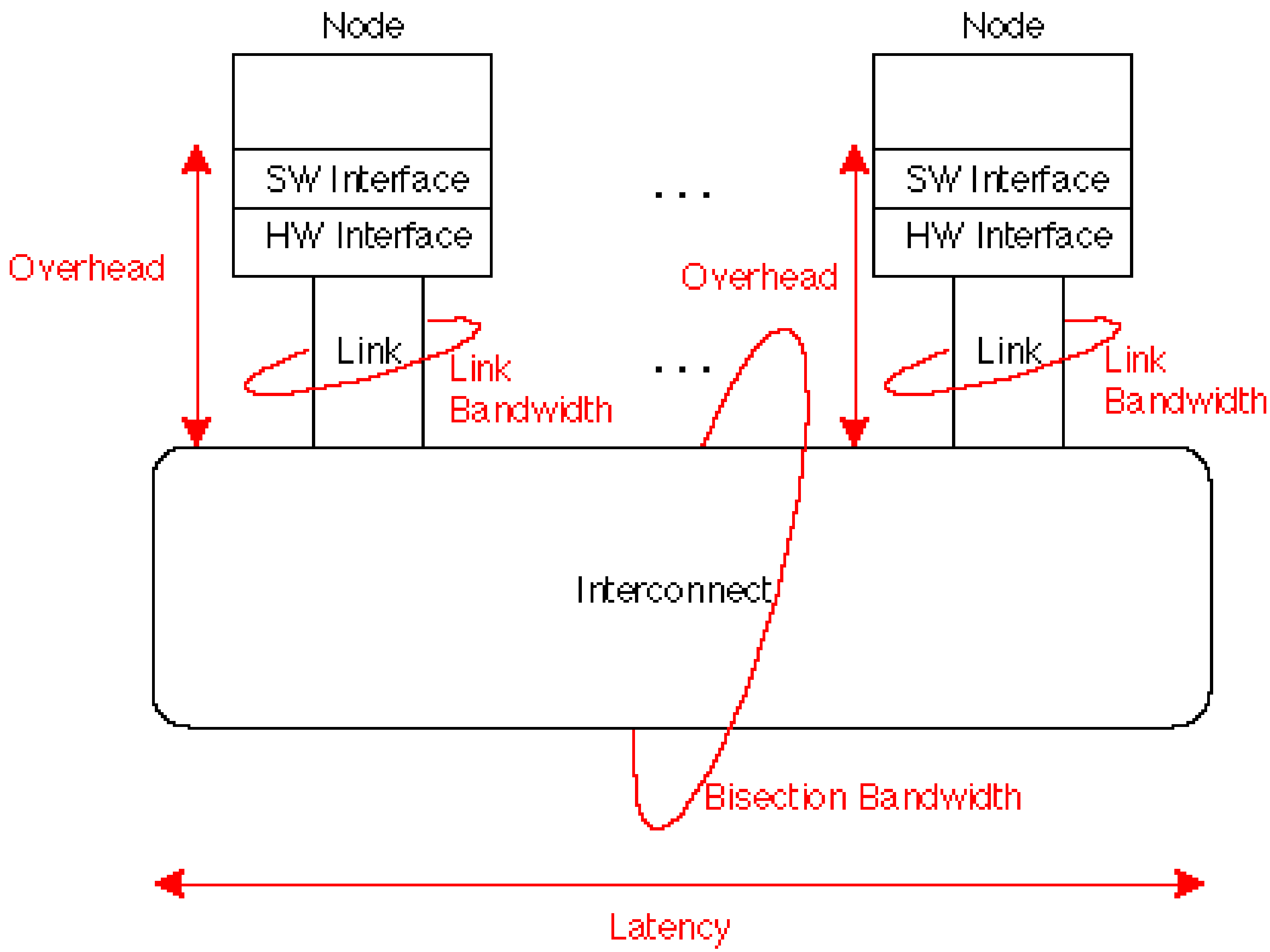


00: Request—Please send data from Address
01: Reply—Packet contains data corresponding to request
10: Acknowledge request
11: Acknowledge reply

Software to Send and Receive

- **SW Send steps**
  - **1: Application copies data to OS buffer**
  - **2: OS calculates checksum, starts timer**
  - **3: OS sends data to network interface HW and says start**
- **SW Receive steps**
  - **3: OS copies data from network interface HW to OS buffer**
  - **2: OS calculates checksum, is matches send ACK; if not, deletes message (sender resends when timer expires)**
  - **1: If OK, OS copies data to user address space and signals application to continue**
- **Sequence of steps for SW: protocol**
  - **Example similar to UDP/IP protocol in UNIX**

Network Performance Measures



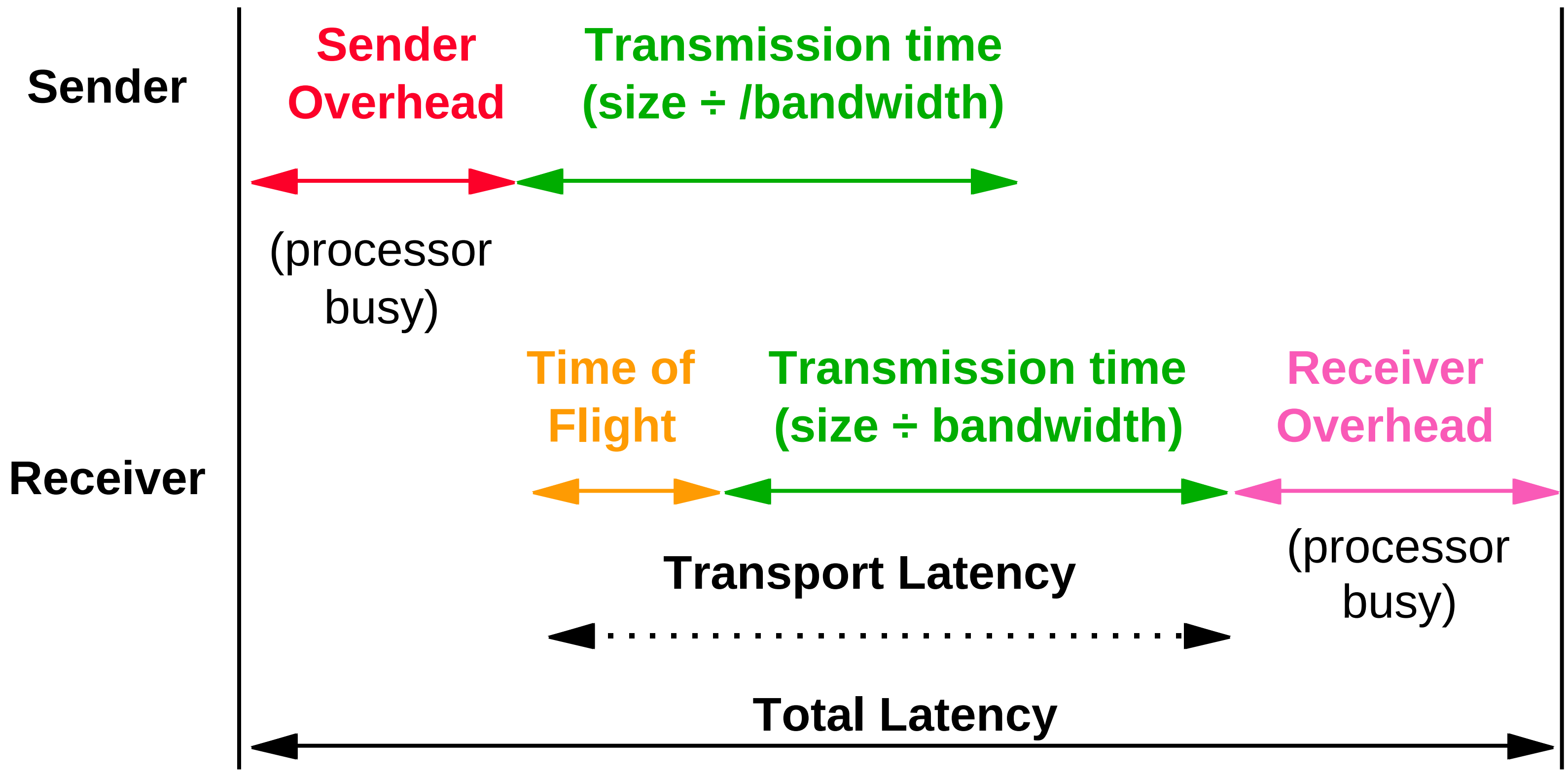
- **Overhead**: **latency of interface vs. Latency**: **network**

Example Performance Measures

Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Bisection BW	N x 5 MB/s	1.125 MB/s	N x 10 MB/s
Int./Link BW	20 MB/s	1.125 MB/s	10 MB/s
Latency	5 µsec	15 µsec	50 to 10,000 µs
HW Overhead to/from	0.5/0.5 µs	6/6 µs	6/6 µs
SW Overhead to/from	1.6/12.4 µs	200/241 µs	207/360 µs

(TCP/IP on LAN/WAN)

Performance Metrics



Total Latency = Sender Overhead + Time of Flight +
Message Size ÷ BW + Receiver Overhead

Total Latency Example

- 10 Mbit/sec., sending overhead of 230 µsec & receiving overhead of 270 µsec.
- a 1000 byte message (including the header), allows 1000 bytes in a single message.
- 2 situations: distance 100 m vs. 1000 km
- Speed of light = 299,792.5 km/sec
- Latency_{100m} =
- Latency_{1000km} =
- Long time of flight => complex WAN protocol

5 minute Class Break

- **Lecture Format:**
  - ≈ 1 minute: review last time & motivate this lecture
  - ≈ 20 minute lecture
  - ≈ 3 minutes: discuss class manangement
  - ≈ 25 minutes: lecture
  - 5 minutes: break
  - ≈25 minutes: lecture
  - ≈1 minute: summary of today’s important topics

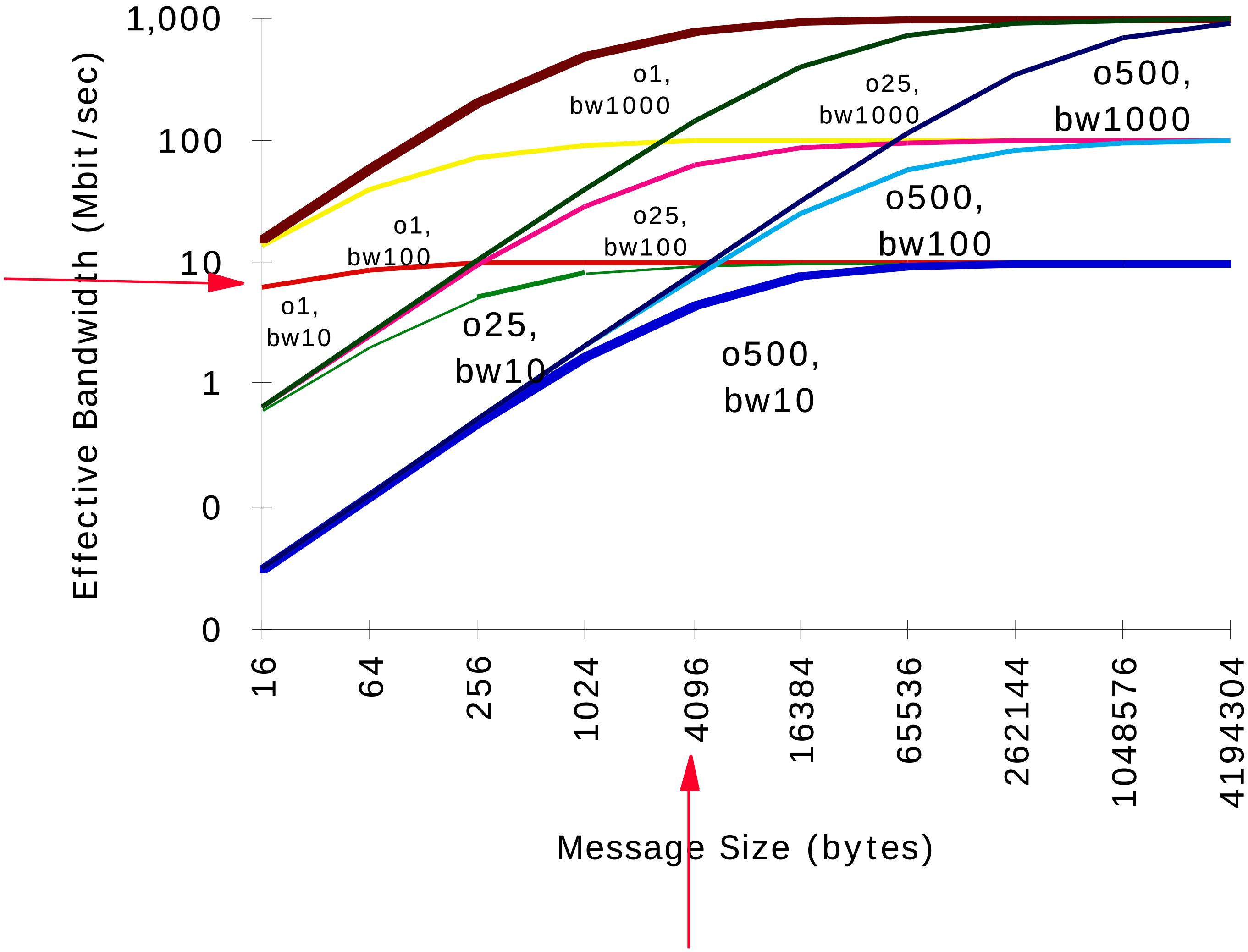
Total Latency Example

- 10 Mbit/sec., sending overhead of 230 µsec & receiving overhead of 270 µsec.
- a 1000 byte message (including the header), allows 1000 bytes in a single message.
- 2 situations: distance 100 m vs. 1000 km
- Speed of light = 299,792.5 km/sec
- $\text{Latency}_{100m} = 230 + 0.1\text{km} / (50\% \times 299{,}792.5) + 1000 \times 8 / 10 + 270$
- $\text{Latency}_{100m} = 230 + 0.67 + 800 + 270 = 1301$ µsec
- $\text{Latency}_{1000km} = 230 + 1000\text{ km} / (50\% \times 299{,}792.5) + 1000 \times 8 / 10 + 270$
- $\text{Latency}_{1000km} = 230 + 6671 + 800 + 270 = 7971$ µsec
- Long time of flight => complex WAN protocol

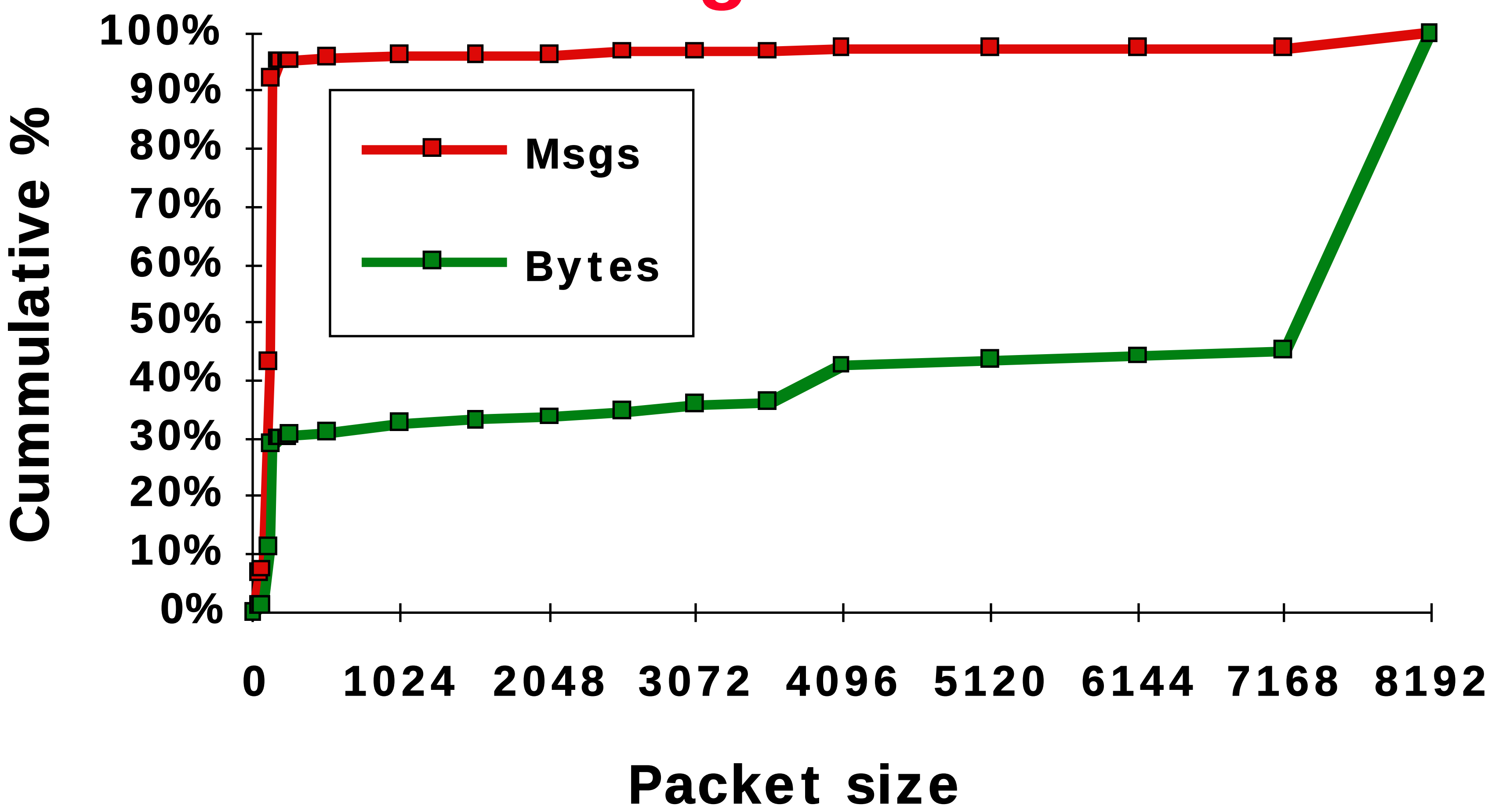
Simplified Latency Model

- **Total Latency ≈ Overhead + Message Size ÷ BW**
- **Overhead = Sender Overhead + Time of Flight + Receiver Overhead**
- **Example: show what happens as vary**
  - **Overhead: 1, 25, 500 µsec**
  - **BW:10,100, 1000 Mbit/sec**
  - **Message Size: 16 Bytes to 4 MB**
- **If overhead 500 µsec, how big a message > 10 Mb/s?**
- **How big are messages?**

Overhead, BW, Size

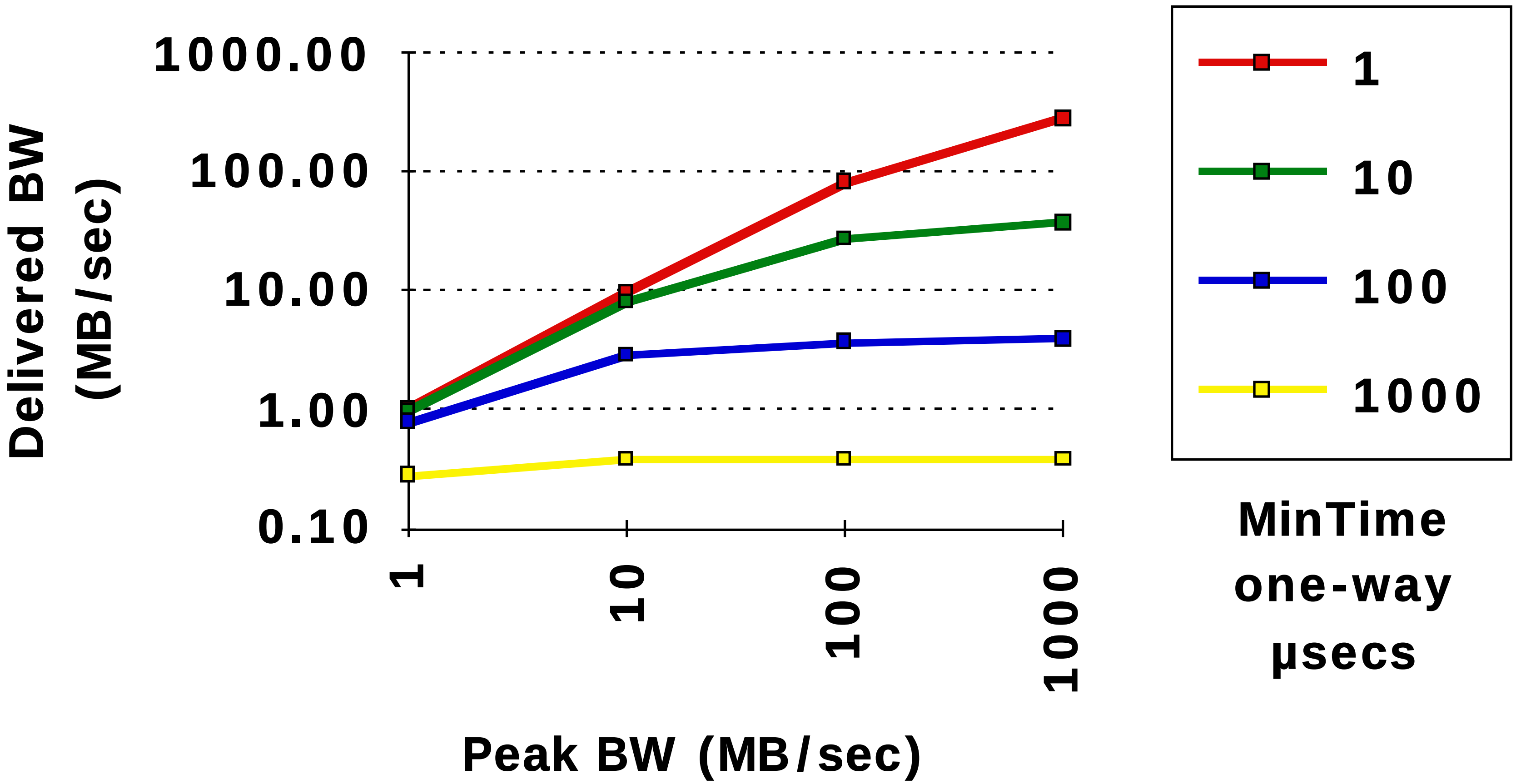


Measurement: Sizes of Message for NFS



- **95% Msgs, 30% bytes for packets ≤ 200 bytes**
- **> 50% data transfered in packets = 8KB**

Impact of Overhead on Delivered BW



- **BW model: Time = overhead + msg size/peak BW**
- **> 50% data transfered in packets = 8KB**

Interconnect Issues

- **Performance Measures**
- **Implementation Issues**
- **Architectual Issues**
- **Practical Issues**

Implementation Issues

Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Maximum length between nodes	25 m	500 m; ≤5 repeaters	copper: 100 m optical: 2 km—25 km
Number data lines	4	1	1
Clock Rate	40 MHz	10 MHz	≥ 155.5 MHz
Shared vs. Switch	Switch	Shared	Switch
Maximum number of nodes	2048	254	> 10,000
Media Material	Copper	Twisted pair copper wire or Coaxial cable	Twisted pair copper wire or optical fiber

Summary: Interconnections

- **Communication between computers**
- **Packets for standards, protocols to cover normal and abnormal events**
- **Performance issues: HW & SW overhead, interconnect latency, bisection BW**
- **Implementation issues: length, width, media**