Lecture #22

****
Review -- 1 min
****

****
Outline - 1 min
****
I/O Motivation and Road Map
Technology Trends
Exam Postmortem
Disk Performance

****
Preview - 1 min
****
Road Map
Storage Systems
    Historical Context of Storage I/O
    Disk Storage Devices
    Performance Measures
    A Little Queuing Theory
    Processor Interface Issues
    I/O Busses
    Tertiary storage
    RAID
    Unix File systems
    I/O Benchmarks
    Comparing Unix File System Performance
Networks

****
Lecture - 20 min
****

Motivation: Why study I/O?
    CPU Performance: 50-100%/year
    Multiprocessor Supercomputers: 150%/year

I/O Limited by *mechanical delays*
    < 10%/year (IO per sec or MB/s)

Amdahl's law: system speedup limited by slowest part
    10% IO & 10x CPU: 5x performance (lose 50%)
    10% IO & 100x CPU: 10x performance (lose 90%)

IO Bottleneck:
    diminishing fraction of time spent in CPU
    →diminishing value of faster cpus

In the end, response time, I/O time is what matters –
    how long from when I hit key until emacs window appears?

Technology trends

CPU: Driven by SPEC – speed, cost/performance
    performance – 55%/year (at constant cost)

Memory: Driven by cost/capacity, commodity market
    performance – 7%/year
    capacity – 4x per 3 years
        (GRAPH)

Disk: driven by cost/capacity
    → 60%-100% per year
    GRAPH

    driven by target markets
    → form factor falls over time (27" → 1.8")

    perf. Limited by mechanical effects
    → 2x per 10years

Disk Historical Perspective

1950's – 1970 – migration from batch to on-line processing
  (big server disks)
  1956 IBM Ramac → early 1970's winchester

  key technology – disk head floats over platters on air buffer
  - 1970's (winchester) – Ics make it cheap enough to put media in sealed box with controller (rather than removable media)
  → head files closer to surface

  steady form factor shrink: 27" → 14"
  driven by
  - performance (higher rotation rate)
  - more actuators in machine room

  story: berkeley 252 field trip to supercomputer center
  they show Illiac disk (early supercomputer)
  theory was: build bigger disk so that transfer rate higher
  for a given rotational speed (edge moves faster)
  made a 5 foot diameter disk
  worry was that if the disk broke, it could come apart and
  people would get killed
  → encase the thing in an inch of steel

1970's –
  5.25" floppy form factor
  (downlod microcode into mainframes to fix ROM mistakes)
  semiconductor memory and microprocessors
  early emergence of industry-standard disk interfaces
  ST506, SASI, SMD, ESDI

1980's
  PCs and first generation workstations
  client/server computing
  accelerates disk downsizing
  8" disk → 5.25"

Mass market disk drives become a reality
industry standards: SCSI, IPI, IDE
5.25” drives for standalone PCs
end of proprietary disk interfaces

Late 1980’s-early 1990s
PCs (31/2”)
Laptops (2.5”)
palmtops/PDAs (1.8”) {kindof a bust}

→ formfactor drives market not performance

optical disk fails on perfomrance (NeXT) but finds niche for software distribution

1990s and beyond– migration to ubiquitous computing
computers in cars, phones, books, thermostat
(thermostat story – company that wants to put thermostat on a network running HTTPD – power company can pay you to turn down your A/C during peak loads)

nationwide fiber optic network with wireless tails

→ ubiquitous computing main goals
smaller
cheaper
lower power

other trend - data utilities
-- high capacity, hierarchical storage (e.g. library of congress)

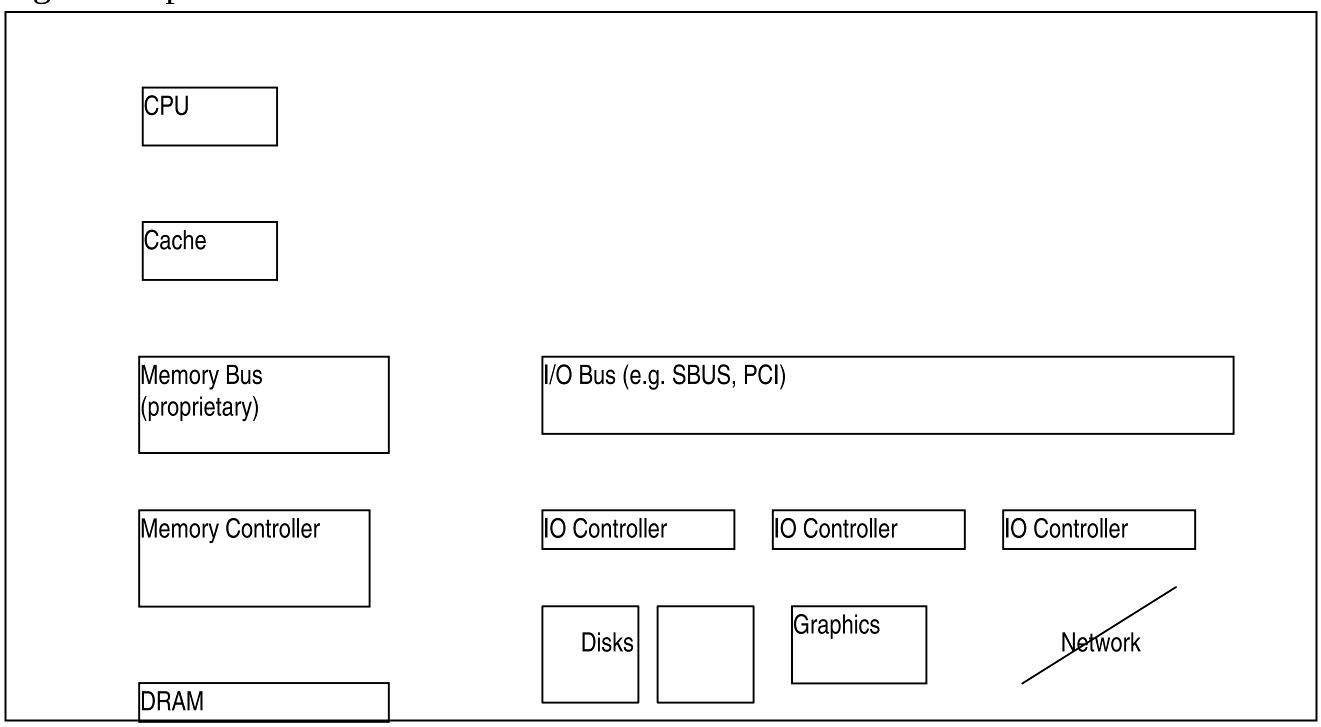
Capacity improvements in the future
quadruples every 3 years (at least)

head flies really close to surface, relies on aerodynamics not to crash
(“like a 747 flying 30’ above sierras”)
(patterson book says 4 microinches)

20 years of future density gains possible using understood techniques

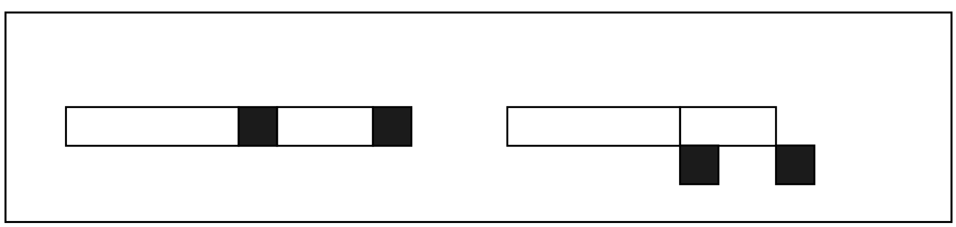
****
Admin - 3 min
****
****
Lecture - 24 min
****

High-level picture of I/O architecture:

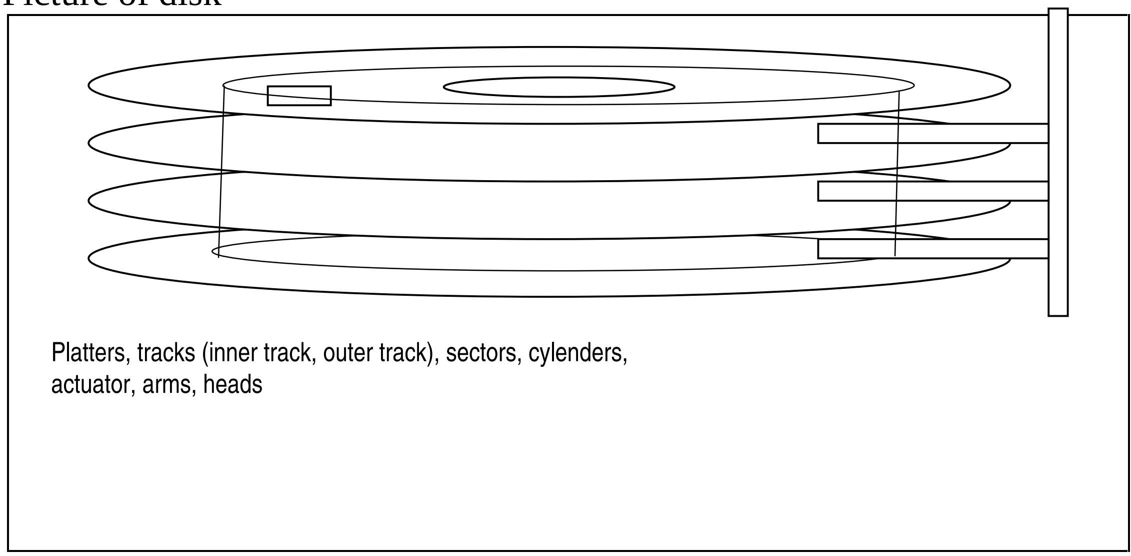


Model for performance

time(workload) = time(CPU) + time(I/O) - time(overlap)



Picture of disk



Density/Capacity improvements
----

example: IBM Ultrastar (top-end disk early 1997)

    9GB
    3.5” form factor
    9 platters (18 recording surfaces)
    829 Mbit/sq.inch
    134,500 bits per inch (max) lineal density
    6160 tracks per inch (track spacing)

trends:

    bpi: increasing slowly – double in 5-10 years
    tpi: most gains come here – 10x in 10 years
    platters – increasess slowly
    overall – expect growth above 60% per year for a while

    sidebar: in early 90’s people were talking about main memory surpassing disk – disk was on a 30-40% curve while memory was on a 60% curve. Memory got expensive for a few years; Disks got scared and have actually increased their lead recently.

disk access time

response time = seek + rot + xfer
+ controller time
+ queue time
+/- cache effects

seek time
----

for back of the envelope:
min seek – 1 sector
max seek – inside to outside
average seek – 1/3 of way across disk (somewhat meaningless)

back of the envelope – real seeks tend to be between min and avg (due to locality)

More details

- better way to look at it - seek profile

short seeks – dominated by settling time (1-3ms)
long seeks – overhead + proportianal to distance

FIGURE 5 – IEEE computer March 1994
<get graph, min, max, avg from IEEE computer>

- Track following

narrower tracks → increased settling time
→ read seek < write seek (speculative reads)

head switch time nontrivial
(due to electronics, *head settling time (tracks no longer precisely aligned across different platters*)
→ "cylinders" less important than they once were

- recalibration
use internal table to decide how far to seek

(store distance to subset of track→ sawtooth pattern)
figure: H&P fig 6.42

thermal expansion, arm pivot-bearing stickiness, etc
→ recalibrate drives internal seek tables
triggered by 1) timer
2) temperature changes
→ happen often during start-up, every 15-30 minutes during steady state

recalibration takes 500+ milliseconds → kills multimeda
(solution – firmware to avoid recalibration or to allow host to schedule recalibration "multimedia disk drives")

Data layout issues
----

zoning – more bytes on outer tracks than inner tracks
(improves capacity, bandwidth)
→ can't just divide address by bytes per track to get track #

sparing – too expensive to manufacture "perfect" surface
→ remap bad sectors
slip sparing – shift numbering by 1
remapping – map bad sector somewhere else on disk

→ makes seek, rotation models complex

Rot time
today 7200RPM is common
→ 8ms per rotation
→ avg wait (after seek): 4ms

subtleties
track skewing

stagger block zero on different tracks to account for seek, head switch

drift in rot. rate

0.5%-1%
(→ hard to model disk rotational position
100-200 revolutions after last known operation)

Transfer time

time for sector to fly under head

e.g. 32 sectors per track
4ms per rotation
→ 1/8 ms to read a track

bandwidth increases more quickly than other factors
(<graph> data from thesis)
reason: capacity drive increases bit per inch
(so not just relying on mechanical)

→ want to transfer large, contiguous units from
e.g. current disk 10MB/s transfer, 10ms seek + rot
→ need to transfer 10ms * 10MB/s = 100 MB to get 50% efficiency
(cylinder probably more like 1MB = 32 sectors * 1KB/sector * 20 tracks/cylinder → 640KB/cyl)

Controller time

nontrivial (see IEEE computer article)

0.3-1ms
(processors get faster at about the same rate that controller programs get more complicated)

Bus transfer time

scsi 5MB/s - 40 MB/s
FIGURE 2: IEEE Computer march 1994
fiber channel

Queue time
next class (or Friday)

Caching
read ahead/read buffer
hard to model (large transfers may bypass cache, replacement policy, etc.)

read ahead really important for sequenial reads (otherwise back-to-back sector reads take 1 full revolution)

write bufer
"immediate reporting" on data writes
(optimize for sequenial transfer broken up by FS)
OS can disable for metadata writes.

Major factor: FIGURE 6 IEEE Computer march 1994

Bottom line
1. tech trends
avg seek time 1987: 16.7ms
1994: 8.0/9.0 ms (read, write)
1999: 5.7/6.5 ms
9%/year

rotational speed/avg latency 1987: 3600rpm/8.3 ms
1994: 7200rpm/4.2 ms
1999: 10000rpm/2.99ms
9%/year

bandwdith: 1987 2.5 MB/s
1994: 6-9 MB/s (inner-outer tracks)

1999: 18-28MB/s
20%/yr

8KB transfer: 1987: 28ms
1994: 13.1 ms
1999: 8.7ms
10%/year

1MB transfer: 1987: 425 ms
1994: 123 ms
1999: 55 ms
(19%/year)

A modern high-performance disk: (March 1999)
Seagate Cheetah:

Capacity	36GB
Internal Transfer rate	18.3-28.0MB/s
External transfer rate	80-100MB/s (Fibre Channel)
Multisegmented cache	1MB
Optional cache	4MB
Track-to-track seek	.6ms(read)/.9ms(write)
Average seek	5.7ms(read)/6.5ms(write)
Maximum seek	12ms(read)/13ms(write)
Average rot latency	2.99ms
Spindle speed	10,000RPM
Disks/heads	12/24
Sectors per drive	71M
Cylinders	9,772

Recoverable read errors per bits read 10 per 10^12
Nonrecoverable read errors per bits read 1 per 10^15

Bottom line 2: disk access timeline (fig 2 IEEE Computer March 1994)

****

Summary - 1 min

****