Overview of the day

▼ Introduction to storage systems
   – storage devices and their workloads
   – request scheduling

▼ Disk arrays
   – high-reliability redundant storage: making sure it’s there when you need it
   – new kinds of disk arrays

▼ Storage area networks
   – connecting storage to its clients
   – CMU’s NASD

▼ Storage management
   – keeping it all together
Acknowledgements

▼ **SSP**: Guillermo Alvarez, Eric Anderson, Ralph Becker-Szendy, Martha Escobar, Susie Go, Michael Hobbs, Kim Keeton, Arif Merchant, Erik Riedel, Cristina Solorzano, Mustafa Uysal, Alistair Veitch

▼ **ex-SSP**: Richard Golding, David Jacobson, Chris Ruemmler, Mirjana Spasojevic

▼ **Others**:
  – Ed Grochowski (IBM Almaden)
  – David Nagle & Garth Gibson (CMU)

▼ **St Andrews University**

▼ **To learn more**:

http://www.hpl.hp.com/SSP/
Introduction to storage systems  [1:15]
Introduction to storage systems

▼ An overview of current storage devices
  – storage hierarchies
  – the storage business

▼ How disk drives work
  – mechanisms
  – technology trends
  – controllers

▼ Request scheduling

▼ Workloads
  – Workload characterization
  – How file systems and databases use storage
Introduction: what are we talking about?

▼ Storage systems
  - the place where persistent data is kept
  - the center of the universe!

▼ Why?
  - information (and hence storage) is key to most endeavours
  - storage is big business (tens of US$b/year)
  - sheer quantities (hundreds of petabytes/year)

  - “Storage will dominate our business in a few years”
    • Compaq VP, 1998
  - “In 3 to 5 years, we will start seeing servers as peripherals to storage”
    • SUN Chief Technology Officer, 1998
Introduction: what are we talking about?

▼ Hardware components
  – Storage devices
    • mechanisms, controllers, packaging
  – Storage connectivity
    • bus/interconnect fabric, host adapters

▼ Software components
  – Critical-path software
    • OS device driver
    • OS logical volume manager
    • File system/database system
  – Storage management software
Introduction: storage hierarchy

- **Primary storage: CPU**
  - registers (1 cycle, a few ns)
  - cache (10-200+ cycles, 0.02–0.5us)
  - local main memory (0.2–4us)
  - NUMA memory (2–10x local memory)

- **Secondary storage ("online storage")**
  - magnetic disks (2–20ms)
  - solid state disks (0.05–0.5ms)
  - cache in storage controllers (0.05–0.5ms)

- **Tertiary storage**
  - removable media: tape cartridges, floppies, CD, … (ms to minutes)
  - tape libraries, optical jukeboxes "nearline" (few s to few minutes)
  - tape vaults (few minutes to days)
Secondary storage devices

▼ Sealed-mechanism magnetic disks ("Winchesters")
   - dominate the industry
   - 1-50+ GB capacity

▼ Other
   - Solid-state disks
     - DRAM package with a battery to look and feel like a disk
   - Promising(?) “new” technologies
     - holographic storage
     - ARS/MEMS (micro-actuators)
     - MRAM (the return of core memory!)
Tertiary storage devices

- Flash-RAM cards (1-100 MB)
- Floppy disks, Iomega zip, removable disk, (1-200 MB)
- CD-ROM, CD-RW (600 MB; replacing floppies in many uses)
- Magneto-optical (MO) disks (0.6-4 GB/platter)
- DVD (up to 4.5 GB; writable DVD “is on its way”)

- Magnetic tapes (1-100 GB/tape)
  - linear format: 1/2” open reel (largely vanished); cartridge tapes
  - helical-scan: DDS (aka “DAT”)
  - serpentine: DLT, Linear Tape Open (LTO)

- Libraries (and vaults)
  - 10-1000+ tapes, CD-ROM, or MO disks
  - pick & load times of few seconds to a couple of minutes
Uses for tertiary storage

▼ Portable, personal data
  – data interchange
  – software distribution

▼ Backups against failures
  – media/site failure
  – user error:
    \[\text{rm * .o; ... "File .o not found"}\]

▼ Archiving for later use
  – ordered, indexed, coherent data
  – banks, credit card companies, insurance
  – life-critical engineering industry (e.g., aircraft engines)

▼ Really big quantities of data
  – NASA satellite data, NSA, ...
1999 DISK/TREND report: revenue projections

Source: DISK/TREND, Inc.
http://www.disktrend.com
June 1999
1999 DISK/TREND report: unit projections

Annual shipments (million units)

Source: DISK/TREND, Inc.
http://www.disktrend.com
June 1999
Hard-disk prices

Source: http://www.pricewatch.com/
16 Feb 2000
Business trend: storage as % of system cost

Source: Gartner group
Business trend: data is moving into databases

WHO OWNS THE DATA?
Data Resident on Open Systems Servers

Source: Systems Research, 1998
IBM Areal Density Perspective
43 Years of Technology Progress

- Travelstar 18GT
- Ultrastar 72ZX
- Deskstar 37 GP

1st GMR Head
1st MR Head
1st Thin Film Head
25% CGR
60% CGR
5 Million X Increase

Areal Density Megabits/in²

RAMAC

Production Year

Ed Grochowski at Almaden
Introduction to storage systems

- An overview of current storage devices
  - storage hierarchies
  - the storage business

- How disk drives work
  - mechanisms
  - technology trends
  - controllers

- Request scheduling

- Workloads
  - Workload characterization
  - How file systems and databases use storage
Magnetic disk drives: what is inside them?

- Host interface
- Speed-matching buffer and cache
- Servo/motor electronics
- Control processor
- R/W electronics
Areal density = linear density * track density

- a disk drive has 1 to 12 platters, 2 heads per platter
- a platter has ~2,000-40,000 tracks
- 1 track contains ~50-200KB
- 1 sector is ~512 B (may be growing to 1-2KB)
Track, Areal, Linear Density Perspective

Track Density, CGR = 30%
Areal Density, CGR = 60%
Linear Density, CGR = 23%

IBM Disk Drive Products
Circles = Server products
Squares = Mobile products

Availability Year

Ed Grochowski at Almaden
Magnetic disk drive: mechanical innards

Source: How hard disks work, Marshall Brain
http://www.howstuffworks.com/hard-disk.htm
1999
Magnetic disk drives: mechanical performance

▼ Seek time
- accelerate (35-40g) [coast]
  slow down
- settle
- single-track seeks
  • “track-switch”
  • special-case performance

▼ Rotational latency
- 3600 RPM ... 5400 ... 7200 ... 10,000 ... 12,000 ...

▼ Head switches
- between platters
- multiple head drives (now extinct)
IBM HDD Access/Seek Time-Performance Increase

Accessing

Seeking

Rotating

seek time \sim \left( \frac{\text{inertia}}{\text{power}} \right)^{1/3} \times \left( \text{data band} \right)^{2/3}

rotational time \sim \left( \text{RPM} \right)^{-1}

(latency)

Ed Grochowski at Almaden
Magnetic disk drives: a few complications

▼ Zoning
  – outer tracks are longer than inner ones
  – tracks have different capacities
    benefits: increased density, higher data rate

▼ Track-skew, cylinder-skew
  – slip the start of the next track by the time it takes to switch to it
    benefit: increased sequential transfer performance

▼ Sparing
  – leave space for when things go wrong; skip over them
Caching
- read-ahead (multiple streams?)
- write behind
- atomicity guarantees (not!)

Controlling the mechanism
- spindle motor
- arm servo-following

Data path management
- DMA control
- protocol sequencing
- request scheduling
Overall I/O performance under real loads

• **Real** is traced I/O load from 1992
• **Simulated** is Chris Ruemmler’s disk simulator (Pantheon progenitor)
• **Demerit figure** is (basically) area between these two curves

HP 97560

Quantum 425
Introduction to storage systems

▼ An overview of current storage devices
  – storage hierarchies
  – the storage business

▼ How disk drives work
  – mechanisms
  – technology trends
  – controllers

▼ Request scheduling

▼ Workloads
  – Workload characterization
  – How file systems and databases use storage
Disk request scheduling

- I/O requests are very bursty
  - queue lengths up to 1000 have been seen
  - especially important for writes

- Queueing takes place in:
  - host device driver
  - disk/array controller
  - in practice: both

- Traditional 1D schemes: minimize seeks
- Better 2D schemes: include rotational latency, too
  - but have to be done in the disk!
1D request scheduling: minimize seeks

- **FCFS/FIFO:** first-in first-out (terrible!)

- **SCAN:** start at one end of the disk, sweep to the other, then reverse direction. **CSCAN:** at end, go straight back to start.

- **[C]LOOK:** like **[C]SCAN**, but go back to first *request*, not start of disk
2D disk request scheduling: min(seek+rotation)

- **Shortest Positioning Time First** (aka Shortest Access Time First)
  - Like cpu scheduling: “do the shortest jobs first”
  - you do well almost all of the time
- **Various age-weighting tricks to avoid starvation**
Scheduling algorithms: performance

HP97560 disk,
8KB reads,
Poisson arrivals;
uniform random distribution
Introduction to storage systems

- An overview of current storage devices
  - storage hierarchies
  - the storage business

- How disk drives work
  - mechanisms
  - technology trends
  - controllers

- Request scheduling

- Workloads
  - Workload characterization
  - How file systems and databases use storage
Workload characterization - why?

- System monitoring
  - What’s going on?

- Improve storage system designs
  - “What if?” design questions
  - Predicting effects of new or “scaled” workloads

- Generate synthetic workloads
  - To test performance of new designs
  - To compare existing systems
Workload characterization: Rubicon (+ Pylon)

- Rubicon is a tool for measuring I/O loads
  - uses HP-UX trace-gathering measurement system
- Pylon is a tool for generating synthetic workloads
  - Rubicon output can be used as Pylon input

- Together ... can test for congruence
  - compare effects of synthetic (replayed) workload against original measurements
Rubicon: sample component tree

- Mux
  - Sends records to multiple receivers
  - Subsets trace records

- Filter
- Analyzer
- Mux
- Analyzer
- Analyzer

Measure stuff; compute workload characterization

Report

Output results in multiple formats

Sends records to multiple receivers

Subsets trace records
Workload characterization: 2 case studies

▼ Electronic mail server
   – HP OpenMail
   – Peak operation period
   – about 1400 active users

▼ Decision support database server
   – Oracle
   – 300 GB TPC-D database
   – Presentation focus: TPC-D Q5
Workload characterization: request size

- Email dominated by small (<= 8 KB) writes
- DSS dominated by larger (64 KB) reads
Workload characterization: fraction of reads

- **Email**
  - average read percent: 28%
  - we need distributions, not just averages
Workload characterization: access locality

- Email
  - Beginning of address range heavily accessed
  - Disk array caching important for performance
Workload characterization: I/O phasing

- Decision support database: TPC-D query
- Request rates vary widely
- Most multi-table queries have multiple phases
Workload characterization: I/O phasing

- Decision support database: TPC-D query
- “Read-only” workload exhibits writes!
Workload characterization: request size

- Decision support database: TPC-D query
- Different behavior from different parts of the database:
  - table, indices, temp space, log
Workload characterization: lessons learned

▼ Lessons learned:
- List of important characteristics is longer than you think
- Distributions, not averages, are important

▼ Characteristics of interest:
- Request size distribution
- Request rate distribution
- Read:write ratio
- Spatial locality (e.g., sequentiality)
- Temporal locality (e.g., data re-references)
- Correlation between accesses to different parts of storage system
- Burstiness
- Phased behavior
Workload characterization: open problems

- Characterizing workloads
  - correlations
  - burstiness (self-similarity at long term)
  - good spatial locality measures

- Replaying workloads
  - accurate timing is the hard part

- Predicting future loads
  - interleaving/workload merging
  - workload scaling
  - modelling application/dbms effects
Summary so far

- Storage devices: disks, tapes, other
- Performance issues: really important!
- Scheduling is way too much fun!
- Application behavior matters!