Disks:
A discussion of:
The Five-Minute Rule Ten Years Later
A Case for Redundant Arrays of Inexpensive Disks
Presented by Matt Becker
(February 28th, 2005)

---

Disks
Presentation Overview

• Preface: A Brief Introduction to Disks.

• The Five Minute Rule Ten Years Later

• Redundant Array of Inexpensive Disks

---

Preface
A Brief Introduction to Disks

• Brief history of disks.
  - What did they use before disks?
  - Why are they so popular now?

• Heads, Platters and Cylinders
  - How does all this stuff work?

• Trends
  - What metrics are important?
  - What trends can we see?
Preface - A Brief History of Disks

• What was used before disks?
  - No storage at all!
  - Paper/Punchcards. (you knew that)
  - Magnetic tape (still used today)

• The first disk drives
  - IBM pioneered most of the work in this area.
  - Heads in contact with the platter
  - A Breakthrough!
  - Lead to the creation of the IBM 305 RAMAC

• Why are they so popular?
  - Nonvolatile, Reliable, Cheap, Random Access, and faster than previous alternatives

---

Preface - Heads, Platters and Cylinders

• How does all of this stuff work?
  - What's inside that little black box?

---
Preface - Heads, Platters and Cylinders

• Heads have come a long way...
  - Used to actually come in contact with the platter surface
  - Wear out and "crash"
  - Advances in heads have helped contribute to the increase in Areal density (more about this later)

Preface - Heads, Platters and Cylinders

• More on platters...
  - The material on the surface.
  - Changes in the composition.
  - Areal Density
• What’s wrong?

Preface - Heads, Platters and Cylinders

• Zoned Bit Recording
Preface - Metrics

• What Metrics are Important?
  - Price
    - Cost/MB of disk space, Cost/Disk Drive
  - Transfer rate
    - Sequential Data Access Rate, Random Access Data Rate
    - Seek Time, Rotational Delay, Transfer Time, Latency, Scan
  - Capacity
    - Capacity/Actuator, Areal Density, Capacity
  - Reliability
    - Mean Time to Failure
  - Physical Specs
    - Number of Platters, Rotational Speed, Number of Actuators, Head Type, Media Type, Encoding, Tracks/Platter, Sectors/Track, Power Consumption
  - Many many more...
  - We are only concerned with a couple!

Preface - Trends

• Capacity over the years...
  - Has been growing very quickly
  - From 10MB in 1981 to 10GB in 2000

• Interesting side-note:
  - XP can't keep up with the increasing size...
Preface - Trends

• A HUGE increase in Areal Density...
  - Grown over 5 Million times since 1960.
  - Mainly due to physical advances (Head/Platter composition).

• Prices are dropping very quickly...

• Seek Time/Transfer Rate
  - Also increasing fairly fast, (as of 2000)
  - The speed at which data can be pulled from the disk is increasing faster though.
  - Leading to the processor memory gap.
  - Hard Disk manufacturers need to address seek time and latency
Preface - Trends

• Other trends that are not so crazy:
  - Reliability
    ▪ Has gone up steadily but is plateauing as manufacturers make disks as good as they can be. Can only make them so reliable.
    ▪ Another reason is that the technology is changing too fast to keep up.
    ▪ The concept of RAID has also made this a less important spec.
  - Interfaces
    ▪ Has changed very slowly.
    ▪ Interfaces are generally the same but with newer standards.
    ▪ Good and bad.
  - Power Consumption
    ▪ Has been decreasing.
  - Internal "Guts"
    ▪ Change more than all the time here. New head designs, new platter materials, better manufacturing practices.
    ▪ Page Size has remained relatively constant despite many projections otherwise. (4KB - 8KB)

Preface - Conclusions

• Introduction to Hard Disks.
  - Before Disks
    ▪ Punch-cards, Floppy Disks, Tape Drives
  - The Breakthrough
    ▪ Head hovers over the Platter surface
  - Basic Hard Disk concepts:
    ▪ Platter, Head, Zone, Cylinder, Track, Sector
    ▪ Seek Time, Rotational Delay, Transfer Time
    ▪ Areal Density, Zoned Bit Recording
  - Metrics and Trends
    ▪ Many metrics, only a few concern us.
      ▪ Cost/MB, Seek Time, Latency, Reliability...
    ▪ Pushing the envelope of what's possible
      ▪ Great increases in Capacity and Areal Density
    ▪ Some things don't change so fast
      ▪ Reliability, Page Size, Interfaces

The Five-Minute Rule

• The Five-Minute Rule
  - What is it? Why is it important?
  - Randomly Accessed Pages
  - Sequentially Accessed Pages
  - Checkpoint strategies
• Index Page Size
  - How big should index pages be?
• New Storage Metrics
The Five-Minute Rule: What is it?

- **What is it?**
  - Disk pages that are referenced every five minutes should be kept in main memory.
  - As disk drives get cheaper the Reference Interval goes up.
  - As RAM gets cheaper the Reference Interval goes up.
  - As the density of pages in RAM goes up the Reference Interval goes down.

  - **Why is it important?**
    - Gives a rule-of-thumb for how long data should remain in main memory.
    - When should you write stuff to a tape? (more on this later)
    - Should we plan on two-phase or one-phase sorts?

The Five-Minute Rule: Randomly Accessed Pages

- **Randomly Accessed Pages**
  - Disk Transfer time is slower
    - Has to account for positioning and reading
    - Usually approximated at 10ms; some modern disks boast <4ms seek times.
    - It's really easy to be fooled because Seek Time can be measured in many ways.
  - **Example in the paper:**
    - PagesPerMBofRAM = 128 (pages/MB) (8K Pages)
    - AccessesPerSecondPerDisk = 64 (accesses/sec/disk) (10ms transfer time)
    - PricePerDiskDrive = 2000 $/disk (9GB + controller)
    - PricePerMBofRAM = 15 $/MB_DRAM
    - ReferenceInterval = 267 seconds

The Five-Minute Rule: Randomly Accessed Pages

- **Problems:**
  - Prices for same equipment vary greatly
    - Server prices are often 3 times the "street" price of the same equipment
  - Any others?
- **Response:**
  - Generally follow between the 1 and 10 min rule.
    - Why is that somewhat ridiculous?
    - Holding technology ratio constant at 2
    - Economic ratio can vary from 30 - 300
    - Holding economic ratio constant at 133
    - Technology ratio can vary from .45 - 4.5
    - Encompasses a vast range of possibilities
    - The difference between 1 and 10 min for a computer is huge
The Five-Minute Rule: Randomly Accessed Pages

- **Modern Example (estimation):**
  - Seagate Barracuda ATA IV 80GB (Slow, cheap)
    - PagesPerMBofRAM = 128
    - AccessesPerSecondPerDisk = 100 (10ms Seek Time)
    - Proc/MBofRAM = 150 (PC100 512KB) expensive
    - ReferenceInterval = 384 (6 min.)
  - Seagate Barracuda SATA 80GB (Really Fast)
    - PagesPerMBofRAM = 128
    - AccessesPerSecondPerDisk = 200 (5ms Seek Time)
    - Proc/MBofRAM = 250 (PC100 512KB) expensive
    - ReferenceInterval = 320 (5 min.)
  - Change the RAM to (DDR PC3200)
    - Proc/MBofRAM = 3 (1 GB PC3200)
    - ATA Reference Interval = 10 min.
    - SATA Reference Interval = 9 min.

The Five-Minute Rule: Sequentially Accessed Pages

- **Operations like Sort, Cube, Rollup, Join have different RAM/Disk tradeoffs**
  - Data is accessed in a sequential manner which greatly improves hard drive throughput.
  - Disk bandwidth drops about 10x when fetching random pages.
  - How does this change the rule original example?
    - Much higher AccessesPerSecondPerDisk (10x higher)
    - Can use different page sizes (let's assume 64KB)
      - PagesPerMBofRAM = 16 (64KB pages)
      - AccessesPerSecondPerDisk = 80 (5 MB/sec)
      - Proc/MBofRAM = 15 (PC100 512KB) expensive
      - ReferenceInterval = 30 seconds
    - Some operations like sort need to read and write data
      - Double the IO cost reference interval becomes more like 1 min.

The Five-Minute Rule: Sequentially Accessed Pages

- **Revisit our modern examples:**
  - Seagate Barracuda ATA IV 80GB (Slow, cheap)
    - PagesPerMBofRAM = 16 (64K pages)
    - AccessesPerSecondPerDisk = 125 (8 MB/sec)
      - (30 pages/sec * 4 MB pages * 500 ms/MB = 8 MB/sec)
    - Proc/MBofRAM = 150 (PC100 512KB) expensive
    - ReferenceInterval = 40 seconds
  - Seagate Barracuda SATA 80GB (Really Fast)
    - PagesPerMBofRAM = 16 (64K pages)
    - AccessesPerSecondPerDisk = 250 (16 MB/sec)
      - (300 pages/sec * 4 MB pages * 500 ms/MB = 8 MB/sec)
    - Proc/MBofRAM = 250 (PC100 512KB) expensive
    - ReferenceInterval = 30 seconds
  - Change the RAM to (DDR PC3200)
    - Proc/MBofRAM = 3 (1 GB PC3200)
    - ATA Reference Interval = 64 seconds
    - SATA Reference Interval = 53 seconds
The Five-Minute Rule: Sequentially Accessed Pages

• Application:
  - A one-pass sorting algorithm is known to run at 5 GB/min.
  - What does the 1 minute rule tell us about this situation?
  - It tells us that for files at or under 5 GB our one-minute rule tells us that everything should fit into RAM.
  - Beyond that we should use two-pass algorithm.

The Five-Minute Rule: Sequentially Accessed Pages

• Application:
  - For sequential access, what happens as the page size grows?
  - Reference Interval asymptotically approaches 40 sec.
  - After page size = 1 MB (1E+6) then the technology ratio becomes the inverse of the disk bandwidth.

The Five-Minute Rule: Checkpoint Strategies

• Let's Consider a Tape drive...
  - From the paper:

| Table 4: Tape robot price and performance characteristics (source Quantum DLTstar™). |
|---------------------------------|---------------------------------|
| Quantum DLT Tape Robot          | 9,000$ price                    |
| Tape capacity                   | 35 GB                           |
| Number of tapes                 | 14                              |
| Robot Capacity                  | 490 GB                          |
| Mount time (rewind, unmount, put, pick, mount, position) | 30 seconds |
| Transfer rate                   | 5 MBps                          |
Two Equations that are important
- \[ \text{IndexHeight} = \log(N) \]
- \[ \text{Utility} = \log(\text{EntriesPerPage}) \]
- IndexHeight is a boilerplate equation. Gives the height of the index for \( N \) pages.
- Utility measures how much closer an index page brings an associated search to the destination record.
- For example, if each index entry is 20 bytes and a 2KB index page that is 70% full contains about 70 entries. This page will have a utility of 6.2. How much for page size: 128KB, 256KB???

Overall Economic Model
- We have utility now we need to have cost...
- Read time of a 2KB index page.
  - 10ms average seek time
  - 10MB/sec. transfer rate
  - How long will it take?
- Suggested Page size is 8KB to 32KB with 16 being the best.

The Five-Minute Rule: Conclusions
- What is it? Why is it important?
  - Tells you how long you should expect to keep pages in memory.
  - Can suggest optimal page sizes.

Randomly Accessed Pages
- Five-Minute rule still applies
- Technology Ratio has gone down ten-fold
- Economic ratio has gone up ten-fold

Sequentially Accessed Pages
- One-Minute rule
- Disks can transfer sequential data faster than randomly accessed data
- Sequential Operations should cache data they expect to revisit within a minute.

Checkpoint Strategies
- Application of same equation used above. Plugging in values you can decide when to archive data to a tape.
  - If the data is not going to be accessed in the next 2 months send it to tape.
The Five-Minute Rule: Conclusions (cont.)

- **Index Page Size**
  - Index Page Benefit / Cost
    - 16 KB is a good size for index pages.

- **New Storage Metrics**
  - Capacity has become unimportant
  - Transfer Speed, Scan Time have become very important

RAID: Redundant Array of Inexpensive Disks

- **The Impending I/O crisis**
  - Growing Gap between Disk and CPU

- **Reliability vs. Performance vs. Cost**
  - Reliability, Performance and Usable Storage Capacity Percentage

- **RAID**
  - What does each RAID level mean?
  - Which one is the best for me?

RAID: The Impending I/O Crisis

- The gap between secondary memory and main memory and CPU is growing.
  - CPU speed has been increasing very quickly
  - Main memory has kept pace
    - Invention of caches and the memory hierarchy
    - SRAM technology used to build cache has improved 40-100% per year.
  - Hard Disks have been improving greatly but not fast enough.
  - To offset the gap manufacturers have just increased the amount of RAM.
    - Helps only for I/O activity with locality of reference
RAID: Reliability vs. Performance

- **Reliability Overhead Cost**
  - The number of extra check disks as a percentage of the total number of disks. (varies from 100% - 4%)

- **Useable Storage Capacity Percentage**
  - The percentage of all the disks (check and otherwise) on which data can be stored. (varies from 50% - 96%)

- **Performance**
  - Number of Reads/Writes per second for blocks of data

---

RAID: Level 0

- Data is broken into blocks and each block is written to different disks.
  - I/O Performance is greatly enhanced.
  - Easy to implement
  - Not really RAID
    - no redundancy, if you lose a disk you can lose a lot of data

---

RAID: Level 1

- Every Disk has a Clone
  - Twice the read rate
  - Same write rate
  - Really high reliability (can sustain multiple drive failures)
  - Very high overhead (100%)
RAID: Level 0+1

- **Mirrored RAID 0 arrays**
  - The performance advantages of RAID 0 with the fault-tolerance of RAID 1
  - Very expensive to implement
  - Not very scalable

RAID: Level 2

- **Hamming Code ECC**
  - Data words are striped across disks by bit and a checksum is computed and stored in a separate disk.
  - Checksum checked on read and updated on write

RAID: Level 3

- **Divide data block and stripe it across disks**
  - Data block is divided and "striped" across disks.
  - Stripe parity is updated on write and checked on read. Kept on a single disk.
  - Optimal Overhead Cost.
  - Small reads and writes suffer because they are spread across disks
RAID: Level 4

- Write entire data block to disk
  - Compute checksum for each rank of blocks updated on write and checked on reads
  - Greatly helps small reads!

RAID: Level 5

- Distribute the check codes through all the disks.
  - In previous levels check disk was a bottleneck.
  - Greatly improved small writes

RAID: Which Level is Best for me?

- Depends on what you are doing?
  - Performance without regard for reliability
    - RAID Level 0 (no redundancy)
  - Performance and reliability but expensive
    - RAID Level 1 or Level 0+1
  - Reading/Writing Large amounts of data at once
    - RAID Levels 2 and 3 are okay
  - Good performance for all types of reads and writes and good reliability
    - RAID Levels 4 and 5
    - RAID Level 5 greatly improves small reads.
RAID: Conclusions

• The Impending I/O Crisis
  - Growing gap between CPU/Main memory and hard disks
  - Hard Disks can't keep up with advances in RAM and CPU technologies
  - Bottleneck on hard disks

• Performance vs. Reliability
  - Reliability Overhead Cost
    - Became a moot point after RAID Level 2
  - Useable Storage Capacity Percentage
    - Again moot point after RAID Level 2.
  - Performance
    - Improvements with each successive level of RAID
      - Early levels (1-2) are okay for large data read/writes
      - Levels 3-5 concentrate on improving performance for small reads and writes.

• What level is best for me?
  - It depends but level 5 is the most versatile