

CSE 451: Operating Systems  
Spring 2006

Module 18  
Redundant Arrays of Inexpensive Disks  
(RAID)

John Zahorjan  
zahorjan@cs.washington.edu  
Allen Center 534

## Managing Physical Resources

- A single disk can be broken up into smaller pieces (partitions):
  - Limit the damage of failures
  - Unit of backup/migration
  - Run different kinds of file systems
- What about making use of multiple disks?
  - None of the file systems we've seen can span multiple disks
    - Why?
  - What about mounting multiple devices?
    - More storage?
    - Better performance?

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

2

## Performance

1. Disk transfer rates are improving, but much less fast than CPU performance
2. We can use multiple disks to improve performance
  - by *striping* files across multiple disks (placing parts of each file on a different disk), we can use parallel I/O to improve access time
3. Striping reduces reliability
  - 10 disks have about 1/10th the MTBF (mean time between failures) of one disk

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

3

## Reliability

- It's typically enough to be resilient to a single disk failure
  - In theory, the odds that another disk fails while you're replacing the first one are low
    - The first time CSE ran a RAID it happened to us...
- To improve reliability, add redundant data to the disks
  - We'll see how in a moment
- So:
  - Performance from striping
  - Resilience from redundancy, which steals back some of the performance gain

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

4

## RAID

- A RAID is a **Redundant Array of Inexpensive Disks**
- Disks are small and cheap, so it's easy to put lots of disks (10s to 100s) in one box for increased storage, performance, and availability
- Data plus some redundant information is striped across the disks in some way
- How striping is done is key to performance and reliability

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

5

## Some RAID tradeoffs

- Granularity
  - fine-grained: stripe each file over all disks
    - high throughput for the file
    - limits transfer to 1 file at a time
  - course-grained: stripe each file over only a few disks
    - limits throughput for 1 file
    - allows concurrent access to multiple files
- Redundancy
  - uniformly distribute redundancy information on disks
    - avoids load-balancing problems
  - concentrate redundancy information on a small number of disks
    - partition the disks into data disks and redundancy disks

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

6

## RAID Level 0: Non-Redundant Striping



- RAID Level 0 is a non-redundant disk array
- Files are striped across disks, no redundant info
- High (single file) read throughput
- Best write throughput (no redundant info to write)
- Any disk failure results in data loss
  - What is lost?

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

7

## RAID Level 1: Mirrored Disks



- Files are striped across half the disks, and mirrored to the other half
  - 2x space expansion
- Reads:
  - Read from either copy
- Writes:
  - Write both copies
- On failure, just use the surviving disk

What is the effect on performance?

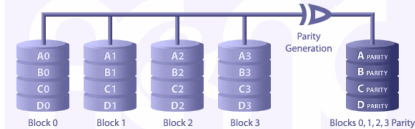
How many simultaneous disk failures can be tolerated?

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

8

## RAID Levels 2, 3, and 4: Striping + Parity Disk



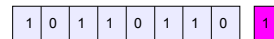
- RAID levels 2, 3, and 4 use ECC (error correcting code) or parity disks
  - E.g., each byte on the parity disk is a parity function of the corresponding bytes on all the other disks
- A large read accesses all the data disks
  - A single block read accesses only one disk (RAID 4)
- A write updates one or more data disks plus the parity disk
- Resilient to single disk failures (How?)
- Better ECC  $\Rightarrow$  higher failure resilience  $\Rightarrow$  more parity disks

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

9

## Refresher: What's parity?



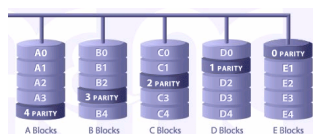
- To each byte, add a bit set so that the total number of 1's is even
- Any single missing bit can be reconstructed
- (Why does memory parity not work quite this way?)
- Think of ECC as just being similar but fancier (more capable)

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

10

## RAID Level 5



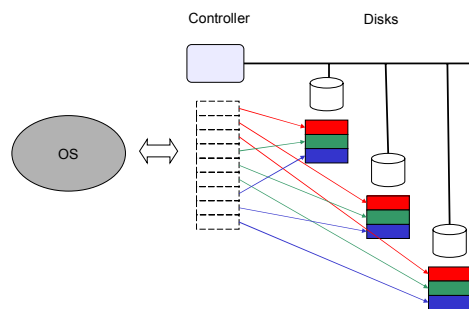
- RAID Level 5 uses block interleaved distributed parity
- Like parity scheme, but distribute the parity info (as well as data) over all disks
  - for each block, one disk holds the parity, and the other disks hold the data
- Significantly better performance
  - parity disk is not a hot spot

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

11

## Typical Implementation



5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

12

## JBOD (Just a Bunch Of Disks)



- Not really a RAID – no striping, no redundancy
- Blocks from disks are concatenated to form a single, logical device
- Allows use of disks of differing sizes

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

13

**PROMISE TECHNOLOGY**  
VTrack 15100 RAID Storage  
**\$5,652.95**  
Usually Ships: 5-7 Days

Cache / Buffer Size: 256 MB  
Data Transfer Rate: Up to 200 MBps (aggregate using both SCSI channels)  
Device Type: RAID Storage System  
Dimensions (WxDxH): 17.0" x 20" x 5" / 65 lbs (without drives)  
Interface Type: SCSI  
Ports Total (Free): 2 x External Ultra160 SCSI (MHDC)  
Power: Dual 500 W; 100-240 VAC auto-ranging, 50-60 Hz, dual hot swp and redundant with PFC, N+1 design  
Power Consumption Operational: 440 Watts (under load)  
RAID Level: RAID 0,1,3,5 or 10 (mirrored stripes), and 50 (striped RAID 5 arrays)  
Channel Qty: 2

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

14

**Dual Channel UltraATA/100 PCI RAID Controller Card**

**\$59.99**  
As low as \$2/month

**SRS Serial ATA-to-Ultra ATA Adapter**  
\$33.95  
**\$27.86** [You Save \$6.10]

Usually Ships: Within 24 Hours

As low as \$1/month

**Overview**

The UltraATA/100 PCI RAID Controller Card has 500MB/sec high-speed data transfer rates up to 100MBps and supports RAID 0 (striping), RAID 1 (mirroring), and RAID 0+1 (mirrored striping) protection. It auto-detects the drive type and the layout to the optimal performance for each connected IDE drive. It conforms to UltraATA/100 specification with full backward support for UltraATA/60, EIDE and ATA-2. IDE hard disk drives. With bus mastering, it reduces I/O processing load on CPUs to increase the system performance. The PCI RAID Controller Card features CRC error-checking which provides data verification and achieves correct data transfer. The ATA software RAID system GUI monitoring utility displays RAID array configuration information (if array sets are configured) as well as adapter and device information for each physical disk.

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

15

## Final Issues

- If you're running a RAID level with sufficient redundancy, do you need backup?
  - What's the difference between RAID and backup?
- Does RAID provide "sufficient" reliability?
  - If you're Amazon?

Tier I	Single path for power and cooling distribution, no redundant components, 99.671% availability.
Tier II	Single path for power and cooling distribution, redundant components, 99.741% availability.
Tier III	Multiple power and cooling distribution paths, but only one path active, redundant components, concurrently maintainable, 99.982% availability.
Tier IV	Multiple active power and cooling distribution paths, redundant components, fault tolerant, 99.995% availability.

5/24/2006

© 2006 Gribble, Lazowska, Levy, Zahorjan

16