

CSE 451: Operating Systems Winter 2007

Module 18 Redundant Arrays of Inexpensive Disks (RAID)

Ed Lazowska
lazowska@cs.washington.edu
Allen Center 570

The challenge

- Disk transfer rates are improving, but much less fast than CPU performance
- 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
- Striping reduces reliability
 - 10 disks have 1/10th the MTBF (mean time between failures) of one disk
- So, we need striping for performance, but we need something to help with reliability / availability

2/20/2007

© 2007 Gribble, Lazowska, Levy, Zahorjan

2

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*
 - Reliability from *redundancy* (which steals back a bit of the performance gain)

2/20/2007

© 2007 Gribble, Lazowska, Levy, Zahorjan

3

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

2/20/2007

© 2007 Gribble, Lazowska, Levy, Zahorjan

4

Some RAID tradeoffs

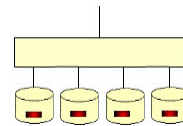
- Granularity
 - fine-grained: stripe each file over all disks
 - high throughput for the file
 - limits transfer to 1 file at a time
 - course-grained: stripe each file over only a few disks
 - limits throughput for 1 file
 - allows concurrent access to multiple files
- Redundancy
 - uniformly distribute 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

2/20/2007

© 2007 Gribble, Lazowska, Levy, Zahorjan

5

RAID Level 0: Non-Redundant Striping



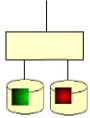
- RAID Level 0 is a non-redundant disk array
- Files are striped across disks, no redundant info
- High (single file) read throughput
- Best write throughput (no redundant info to write)
- Any disk failure results in data loss
 - What is lost?

2/20/2007

© 2007 Gribble, Lazowska, Levy, Zahorjan

6

RAID Level 1: Mirrored Disks



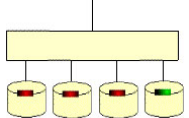
- Files are striped across half the disks, and mirrored to the other half
 - 2x space expansion
- Reads:
 - Read from either copy
- Writes:
 - Write both copies
- On failure, just use the surviving disk

What is the effect on performance?

How many simultaneous disk failures can be tolerated?

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 7


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 ⇒ higher failure resilience ⇒ more parity disks

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 8

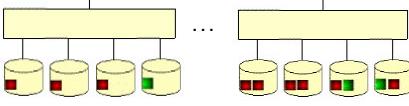
Refresher: What's parity?



- 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)

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 9

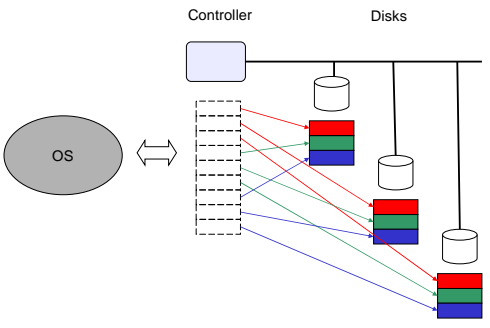
RAID Level 5



- 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


2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 10

Typical Implementation



2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 11

RAID 0-1



Overview
 The UltraATA/100 PCI RAID Controller Card from 3DR achieves high-speed data transfer rates up to 130MB/s and supports RAID 0 (striping), RAID 1 (mirroring), and RAID 0+1 (mirro-striping) protection. It auto-detects the drive type and line rates to the optimal performance for each connected IDE drive. It conforms to UltraATA/100 specification with full backward support for UltraATA/60, IDE and ATA-2 IDE hard disk drives. With hot-swapping, it reduces I/O processing load on CPU to increase the system performance. The PCI RAID Controller Card features ONI 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.

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 12

Buy Online or Contact Us: 1.888.221.2205

RAID 5

Dell recommends Windows Vista™ Business.

You are here: USA > Small Business > Accessories > Storage, Drives & Media > Controller Cards > EE | ATA | SATA

9500S-4LP StorSwitch Serial ATA RAID Controller

Product Details

~~\$439.95~~
\$368.95
(You Save 16%)

Usually Ships Within 24 Hours

ADD TO CART

JWARE

Be the first to write a review

Overview

The 9500S-4LP StorSwitch™ Serial ATA RAID Controller from Jware® features StorSwitch™ enhanced RAID architecture that unloads the power of Serial ATA (SATA), delivering high hardware RAID performance. The 9500S hardware RAID controller delivers in excess of 400 MB per second (MB/sec) sustained RAID 5 reads and over 150 MB/sec RAID 5 sequential writes with less than 2% CPU utilization. The StorSwitch™ architecture can simultaneously handle commands to and from multiple drives. No single drive is penalized by another while executing read or write commands. The on-board processor ensures the host CPU is managing your applications, not the RAID controller hardware. The result is excellent overall performance compared. Applications that benefit most from these high performance controllers include file storage, web servers, cluster servers, supercomputing, real-time backup and archival, security systems as well as streaming applications for audio and video servers.

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 13

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.com?

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

2/20/2007 © 2007 Gribble, Lazowska, Levy, Zahorjan 14