CSE 451: Operating Systems
Autumn 2010

Module 19
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

Reliability

- At the scales we're currently considering (tens of disks), it's typically enough to be resilient to the failure of a single disk
  - What are the chances that a second disk will fail before you've replaced the first one?
    - Er, it has happened to us!
- To achieve this level of reliability, add redundant data that allows a single disk failure to be tolerated
  - We'll see how in a minute
- So:
  - Obtain performance from striping
  - Obtain reliability from redundancy (which steals back some of the performance gain)

RAID

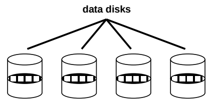
- A RAID is a Redundant Array of Inexpensive Disks
- Disks are small and cheap, so it's easy to put lots of disks (10s, say) in one box for increased storage, performance, and availability
- Data plus some redundant information is striped across the disks in some way
- How striping is done is key to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

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

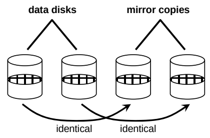
RAID Level 0: Non-Redundant Striping

- RAID Level 0 is a <u>non-redundant</u> disk array
- Files/blocks 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



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



11/21/2010

Prelude to RAID Levels 2-5: A parity refresher

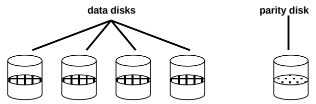
1	0	1	1	0	1	1	0		1

- To each byte, add a bit whose value is 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?)
- More sophisticated schemes (e.g., based on Hamming codes) can correct multiple bit errors – called ECC (error correcting codes)

11/21/2010

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

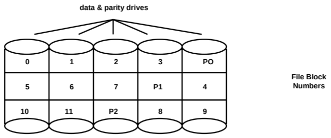
- RAID levels 2, 3, and 4 use parity or ECC disks
  - e.g., each byte on the parity disk is a parity function of the corresponding bytes on all the other disks
  - details between the different levels have to do with kind of ECC used, and whether it is bit-level or block-level
- A read accesses all the data disks, a write accesses all the data disks plus the parity disk
- On disk failure, read the remaining disks plus the parity disk to compute the missing data



11/21/2010

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



0	1	2	3	PO
5	6	7	P1	4
10	11	P2	8	9

11/21/2010

RAID Level 6

- Basically like RAID 5 but with replicated parity blocks so that it can survive two disk failures.
- Useful for larger disk arrays where multiple failures are more likely.

11/21/2010

Example RAID Storage

Promise 3U rack-mountable 16-disk RAID Storage System

Hot swappable drives

Dual controllers with 4 host interface ports for reliability

Can be ganged together into larger units

11/21/2010