

CS 537

Lecture 16

Reliable Storage

Michael Swift

1

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
 - 100 disks have 1/100th the MTBF (mean time between failures) of one disk
- So, we need striping for performance, but we need something to help with reliability / availability
- To improve reliability, we can add redundant data to the disks, in addition to striping

4/9/09

© 2005 Gribble, Lazowska, Levy

2

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

4/9/09

© 2005 Gribble, Lazowska, Levy

3

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

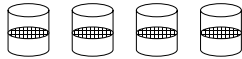
4/9/09

© 2005 Gribble, Lazowska, Levy

4

RAID Level 0

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



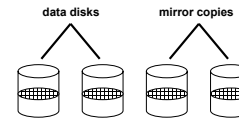
4/9/09

© 2005 Gribble, Lazowska, Levy

5

RAID Level 1

- RAID Level 1 is mirrored disks
- Files are striped across half the disks
- Data is written to two places – data disks and mirror disks
- On failure, just use the surviving disk
- 2x space expansion



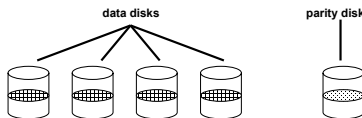
4/9/09

© 2005 Gribble, Lazowska, Levy

6

RAID Levels 2, 3, and 4

- 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 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



4/9/09

© 2005 Gribble, Lazowska, Levy

7

Refresher: What's parity?

1 0 1 1 0 1 1 0 1

- 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?)

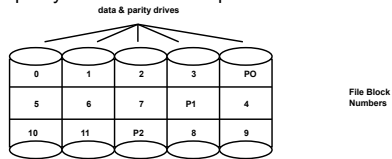
4/9/09

© 2005 Gribble, Lazowska, Levy

8

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



4/9/09

© 2005 Gribble, Lazowska, Levy

9

PROMISE TECHNOLOGY
VTrak 15100 RAID Storage
\$5,652.95
Usually Ships: 5-7 Days
[Add to Orderform](#)

Cache / Buffer Size: 256 MB
Data Transfer Rate: Up to 200 MBps (aggregate using both SCSI channels)
Device Type: RAID Storage System
Dimensions (WxDxH) / Weight: 17.6" x 26" x 5" / 65 lbs (without drives)
Interface Type: SCSI
Port(s) Total (Free) / Connector Type: 2 x External Ultra160 SCSI (M-DCC)
Power: Dual 500 W, 100-240 VAC auto-ranging, 50-60 Hz, dual hot swap 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

10