RAID Disk Arrays

Hank Levy

Basic Problems

- **Disks are improving, but much less fast than CPUs**
- **We can use multiple disks for improving 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**

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 that striping is done is key to performance and reliability.**

Some Raid Issues

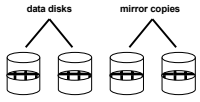
- **Granularity**
  - fine-grained: stripe each file over all disks. This gives high thruput for the file, but limits to transfer of 1 file at a time
  - course-grained: stripe each file over only a few disks. This limits thruput for 1 file but allows more parallel file access
- **Redundancy**
  - uniformly distribute redundancy info on disks: avoids load-balancing problems
  - concentrate redundancy info on a small number of disks: partition the set into data disks and redundant disks

Raid Level 0

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

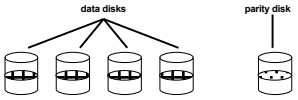
Raid Level 1

- **Mirrored Disks**
- **Data is written to two places**
- **On failure, just use surviving disk**
- **On read, choose fastest to read**



Raid Levels 2 and 3

- 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 data disks plus the parity disk
- On disk failure, read remaining disks plus parity disk to compute the missing data



Level 5

- Block Interleaved Distributed Parity
- Like parity scheme, but distribute the parity info over all disks (as well as data over all disks)
- Better read performance, large write performance