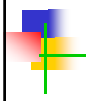


Cluster Computing

Introduction to Software RAID and Parallel Filesystems

- Introduction
- Physical Placement of Data

Introduction

- I/O problems
  - Cluster of workstations has the great amount of resources
  - These large number of resources were not accessible from all the nodes in the network
  - Only the process running on the node that had a given resource attached were able to use it
  - If there was a way to access those remote resources, the steps that had to be done were neither simple nor transparent
  - These resources have to be accessible from any node (Single System Image)

Introduction

- I/O problems
  - These cluster of workstations end up being very similar to parallel machines
    - The same kind of applications can be run on them
    - The problem is founded when executing these applications
      - need a high performance I/O system
    - These applications work with very large data sets, which cannot be kept in memory
    - They expect a fast file system that is able to write & read this data very rapidly
    - When parallel applications are run on the cluster of workstations, the I/O system should allow cooperative operations

Introduction

- I/O problems
  - Need to design high performance file systems
    - Simplify the cooperation between processes
    - Be able to use all the resources efficiently & in an a transparent way

Introduction

- Using Clusters to Increase the I/O performance
  - To achieve a high performance file system
    - need to examine the characteristics a cluster of workstations has & the way we should use them to build better file system
  - Advantages
    - Great quantity of resources
      - Disks that can be used in parallel
      - Large amounts of memory to build big filesystem caches
    - High-speed interconnection network
      - Relay on remote nodes to perform many tasks
      - Use the memory of a remote node for cache blocks
    - Getting closer to parallel machines
      - Parallel machine's skills can be applied

Physical Placement of Data

- Designing a file system for a cluster of workstations
  - Problems
    - Visibility
      - On one hand, many disks scattered among the nodes
      - On the other hand, want to use them from any node in the cluster
    - Achieving a high performance I/O system
      - Disks are mainly built of mechanical components that slow down the most common operations (head movement, disk rotation, etc)
  - The only solution left to increase the disk performance at this level
    - Placing the data in such a way that the mechanical parts have as little effect as possible on the global disk performance

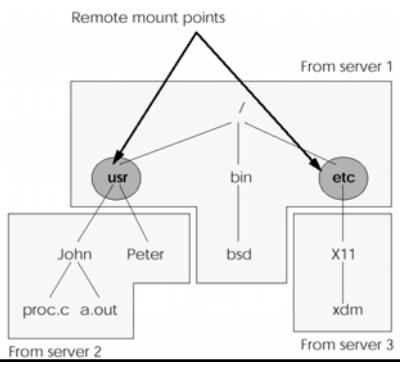
7

Physical Placement of Data

- Increasing the Visibility of the File systems
  - The first problem found in a cluster of workstations
    - Small visibility
      - While many disks are available, only the ones attached to the node where a process is running are visible to that process
    - Many distributed systems have used the mount concept of Unix to increase the visibility
  - Mounting Remote Filesystems
  - To maintain the remote-mount information
    - Two possibility
      - maintaining the mount information at clients (NFS)
      - to maintain the mount information at servers (Sprite filesystem)
    - Solution: caching mechanism

8

Directory Tree Build Mixing Remote and Local Filesystems



9

Physical Placement of Data

- Name Resolution
  - This consists of locating a file or directory given its name
  - Two approaches to this problem
    - Centralized name-resolution scheme
    - Distributed name-resolution scheme
  - Centralized name-resolution scheme
    - One node is responsible for mapping table
    - A failure in the node results in a failure of the whole filesystem
    - Centralized server might become a bottleneck in larger systems

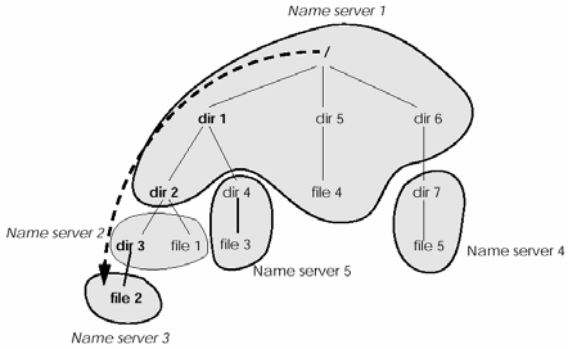
10

Physical Placement of Data

- Name Resolution
  - Distributed name-resolution scheme
    - Two different ways
      - Each system builds its own name space (Sun NFS)
        - Each system knows the filesystems that have been mounted and the node that holds them - movement problem
      - There is a unique global structure (domain)
        - single name space for all workstations
        - name sever is responsible for one of these domains
        - to increase the performance of this name resolution, the system may keep a cache of which nodes have the most popular files or directories

11

Example of Dividing a Directory Tree into Domains



12

Physical Placement of Data

- Data Striping
  - Distribute the data among the disks so that it can be fetched from as many disks as possible in parallel
  - The first time this idea was used was in building a high-bandwidth "single disk"
    - Connect several disks to a single controller and give the impression that the disk had a higher data transfer bandwidth
  - RAID – Redundant Arrays of Inexpensive Disks

13

Physical Placement of Data

- RAIDs
  - Three reasons of the high performance
    - Data from each disk can be fetched at the same time, increasing the disk bandwidth
    - All disks can perform the seek operation in parallel, decreasing its times
    - More than one request may be handled in parallel
  - Data interleaving
    - Fine-grained disk arrays
      - Interleave data in relatively small units so that all I/O requests access all the disks in the disk array
      - Very high data transfer rates for all I/O requests
      - Only one logical I/O request can be served
      - Waste time positioning for every request
    - Coarse-grained disk arrays
      - Interleave data in relatively large units so that all I/O requests need only to access a small number of disks & large requests can access all disks
      - Multiple small requests to be serviced in parallel
  - RAID needs a fault-tolerance mechanism to allow a disk failure without losing the information kept in the failed disk

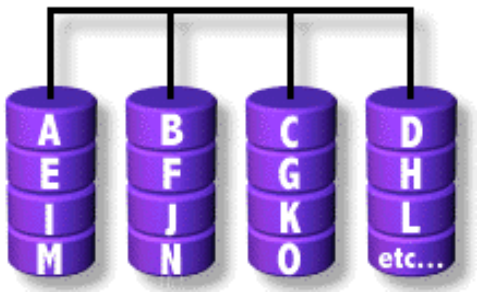
14

RAID

- RAID Level 0
  - Characteristics/Advantages
    - RAID 0 implements a striped disk array, the data is broken down into blocks and each block is written to a separate disk drive
    - I/O performance is greatly improved by spreading the I/O load across many channels and drives
    - Best performance is achieved when data is striped across multiple controllers with only one drive per controller
    - No parity calculation overhead is involved
    - Very simple design
    - Easy to implement
  - Disadvantages
    - Not a "True" RAID because it is NOT fault-tolerant
    - The failure of just one drive will result in all data in an array being lost
    - Should never be used in mission critical environments
  - Recommended Applications
    - Video Production and Editing
    - Image Editing
    - Pre-Press Applications
    - Any application requiring high bandwidth

15

RAID 0: Striped Disk Array without Fault Tolerance



RAID Level 0 requires a minimum of 2 drives to implement

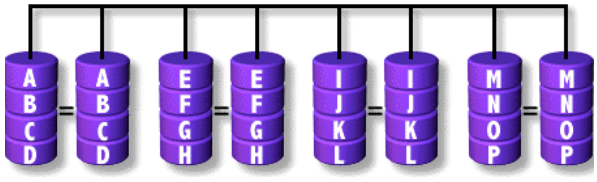
16

RAID

- RAID Level 1
  - Characteristics/Advantages
    - One Write or two Reads possible per mirrored pair
    - Twice the Read transaction rate of single disks, same Write transaction rate as single disks
    - 100% redundancy of data means no rebuild is necessary in case of a disk failure, just a copy to the replacement disk
    - Transfer rate per block is equal to that of a single disk
    - Under certain circumstances, RAID 1 can sustain multiple simultaneous drive failures
    - Simplest RAID storage subsystem design
  - Disadvantages
    - Highest disk overhead of all RAID types (100%) - inefficient
    - Typically the RAID function is done by system software, loading the CPU/Server and possibly degrading throughput at high activity levels. Hardware implementation is strongly recommended
    - May not support hot swap of failed disk when implemented in "software"
  - Recommended Applications
    - Accounting
    - Payroll
    - Financial
    - Any application requiring very high availability

17

RAID 1: Mirroring and Duplexing



For Highest performance, the controller must be able to perform two concurrent separate Reads per mirrored pair or two duplicate Writes per mirrored pair.

RAID Level 1 requires a minimum of 2 drives to implement

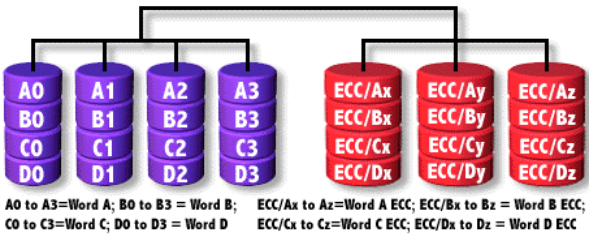
18

RAID

- RAID Level 2
  - Characteristics/Advantages
    - "On the fly" data error correction
    - Extremely high data transfer rates possible
    - The higher the data transfer rate required, the better the ratio of data disks to ECC disks
    - Relatively simple controller design compared to RAID levels 3,4 & 5
  - Disadvantages
    - Very high ratio of ECC disks to data disks with smaller word sizes - inefficient
    - Entry level cost very high - requires very high transfer rate requirement to justify
    - Transaction rate is equal to that of a single disk at best (with spindle synchronization)
    - No commercial implementations exist / not commercially viable

19

RAID 2: Hamming Code ECC



A0 to A3=Word A; B0 to B3 = Word B; C0 to C3=Word C; D0 to D3 = Word D

ECC/Ax to Az=Word A ECC; ECC/Bx to Bz = Word B ECC; ECC/Cx to Cz=Word C ECC; ECC/Dx to Dz = Word D ECC

Each bit of data word is written to a data disk drive (4 in this example: 0 to 3).
Each data word has its Hamming Code ECC word recorded on the ECC disks.
On Read, the ECC code verifies correct data or corrects single disk errors

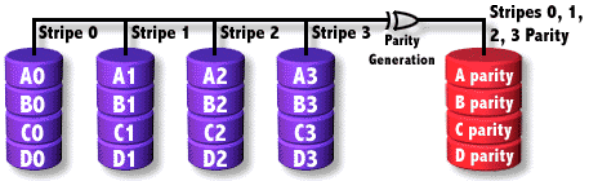
20

RAID

- RAID Level 3
  - Characteristics/Advantages
    - Very high read data transfer rate
    - Very high write data transfer rate
    - Disk failure has an insignificant impact on throughput
    - Low ratio of ECC (Parity) disks to data disks means high efficiency
  - Disadvantages
    - Transaction rate equal to that of a single disk drive at best (if spindles are synchronized)
    - Controller design is fairly complex
    - Very difficult and resource intensive to do as a "software" RAID
  - Recommended Applications
    - Video production and live streaming
    - Image editing
    - Video editing
    - Prepress applications
    - Any application requiring high throughput

21

RAID 3: Parallel Transfer with Parity



The data block is subdivided ("striped") and written on the data disks.
Stripe parity is generated on Writes, recorded on the parity disk and checked on Reads.

RAID Level 3 requires a minimum of 3 drives to implement

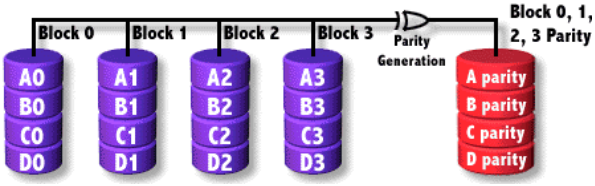
22

RAID

- RAID Level 4
  - Characteristics/Advantages
    - Very high read data transaction rate
    - Low ratio of ECC (Parity) disks to data disks means high efficiency
    - High aggregate read transfer rate
  - Disadvantages
    - Quite complex controller design
    - Worst write transaction rate and write aggregate transfer rate
    - Difficult and inefficient data rebuild in the event of disk failure
    - Block read transfer rate equal to that of a single disk

23

RAID 4: Independent Data Disks with Shared Parity Disk



Each entire block is written onto a data disk. Parity for same rank blocks is generated on writes, recorded on the parity disk and checked on Reads.

RAID Level 4 requires a minimum of 3 drives to implement

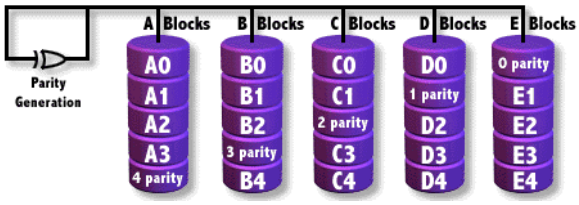
24

RAID

- RAID Level 5
  - Characteristics/Advantages
    - Highest read data transaction rate
    - Medium write data transaction rate
    - Low ratio of ECC (Parity) disks to data disks means high efficiency
    - Good aggregate transfer rate
  - Disadvantages
    - Disk failure has a medium impact on throughput
    - Most complex controller design
    - Difficult to rebuild in the event of a disk failure (as compared to RAID level 1)
    - Individual block data transfer rate same as single disk
  - Recommended Applications
    - File and application servers
    - Database servers
    - WWW, E-mail, and News servers
    - Intranet servers
    - Most versatile RAID level

25

RAID 5: Independent Data Disks with Distributed Parity Blocks

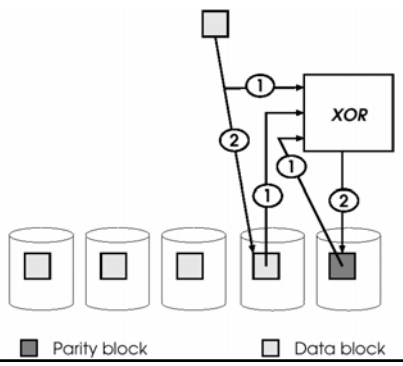


Each entire data block is written on a data disk; parity for blocks in the same rank is generated on Writes, recorded in a distributed location and checked on Reads.

RAID Level 5 requires a minimum of 3 drives to implement

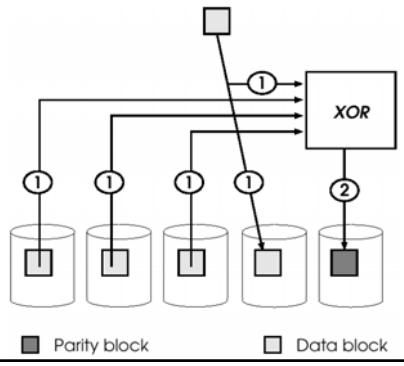
26

Read-Write-Modify



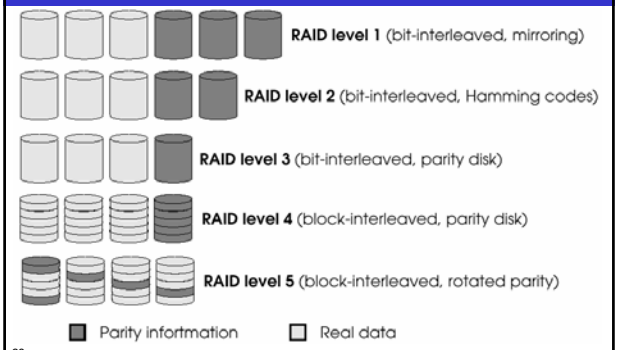
27

Regenerate-Write



28

Graphic Representation of the Five Levels of RAIDs



29

Comparison Between RAID Levels

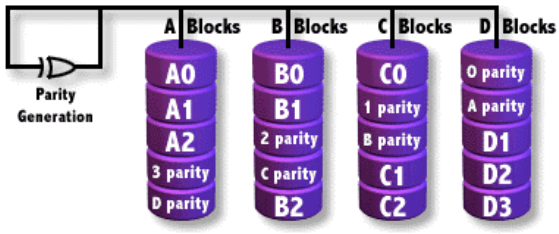
RAID Level	Data Reliability	Data Transfer Rate	I/O Request Rate	Application Strength	Fault-Tolerance Overhead
1	It can handle multiple disk failures in many cases.	Data transfer can be done in parallel, but only half of the disks can be used.	Read: two operations can always be performed in parallel (at least). Write: one operation at a time.	Any	Redundancy uses half the numbers of available disks
2	It can handle multiple disk failures in many cases.	N - (log N) disks can be used in parallel. Codes have to be computed by hardware.	Similar to twice that of a single disk.	Any	Redundancy uses multiple disks (log n).
3	Higher than a single disk. It can handle the failure of one disk.	Highest of all types listed here for reading and writing.	All disks are always used in parallel (byte interleaved).	Multimedia (Video, imaging, sound, ...) Applications that use large files.	Only one disk is used for redundancy.
4	Higher than a single disk. It can handle the failure of one disk.	Reads use all disks but one in parallel. Writes may become slow for parity reasons.	Read: significantly high due to the transfer parallelism. Write: slow when small writes are needed.	Applications with many reads and few writes. Application with large writes.	Only one disk is used for redundancy.
5	Higher than a single disk. It can handle the failure of one disk.	Reads use all disks but one in parallel. Writes may become slow for parity reasons.	Read: significantly high due to the transfer parallelism. Write: slow when small writes are needed.	Transaction processing where many more reads than writes are done.	Only one disk is used for redundancy.

RAID

- RAID Level 6
  - Characteristics/Advantages
    - An extension of RAID level 5 which allows for additional fault tolerance by using a second independent distributed parity scheme (two-dimensional parity)
    - Data is striped on a block level across a set of drives, just like in RAID 5, and a second set of parity is calculated and written across all the drives
    - RAID 6 provides for an extremely high data fault tolerance and can sustain multiple simultaneous drive failures
    - Perfect solution for mission critical applications
  - Disadvantages
    - Very complex controller design
    - Controller overhead to compute parity addresses is extremely high
    - Very poor write performance
    - Requires N+2 drives to implement because of two-dimensional parity scheme

31

RAID 6: Independent Data Disks with Two Independent Distributed Parity Schemes



32

RAID

- RAID Level 7
  - Architectural Features
    - All I/O transfers are asynchronous, independently controlled and cached including host interface transfers
    - All reads and write are centrally cached via the high speed x-bus
    - Dedicated parity drive can be on any channel
    - Fully implemented process oriented real time operating system resident on embedded array control microprocessor
    - Embedded real time operating system controlled communications channel
    - Open system uses standard SCSI drives, standard PC buses, motherboards and memory SIMMs
    - High speed internal cache data transfer bus (X-bus)
    - Parity generation integrated into cache
    - Multiple attached drive devices can be declared hot standbys
    - Manageability: SNMP agent allows for remote monitoring and management

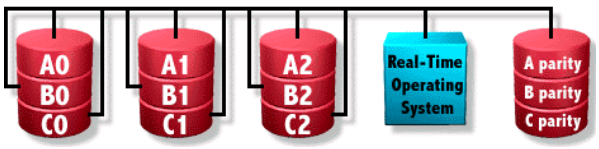
33

RAID

- RAID Level 7 (cont'd)
  - Characteristics/Advantages
    - Overall write performance is 25% to 90% better than single spindle performance and 1.5 to 6 times better than other array levels
    - Host interfaces are scalable for connectivity or increased host transfer bandwidth
    - Small reads in multi user environment have very high cache hit rate resulting in near zero access times
    - Write performance improves with an increase in the number of drives in the array
    - Access times decrease with each increase in the number of actuators in the array
    - No extra data transfers required for parity manipulation
    - RAID 7 is a registered trademark of Storage Computer Corporation.
  - Disadvantages
    - One vendor proprietary solution
    - Extremely high cost per MB
    - Very short warranty
    - Not user serviceable
    - Power supply must be UPS to prevent loss of cache data

34

RAID 7: Optimized Asynchrony for High I/O Rates as well as High Data Transfer Rates



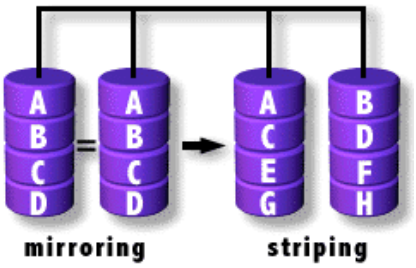
35

RAID

- RAID Level 10
  - Characteristics/Advantages
    - RAID 10 is implemented as a striped array whose segments are RAID 1 arrays
    - RAID 10 has the same fault tolerance as RAID level 1
    - RAID 10 has the same overhead for fault-tolerance as mirroring alone
    - High I/O rates are achieved by striping RAID 1 segments
    - Under certain circumstances, RAID 10 array can sustain multiple simultaneous drive failures
    - Excellent solution for sites that would have otherwise gone with RAID 1 but need some additional performance boost
  - Disadvantages
    - Very expensive / High overhead
    - All drives must move in parallel to proper track lowering sustained performance
    - Very limited scalability at a very high inherent cost
  - Recommended Applications
    - Database server requiring high performance and fault tolerance

36

RAID 10: Very High Reliability Combined with High Performance



RAID Level 10 requires a minimum of 4 drives to implement

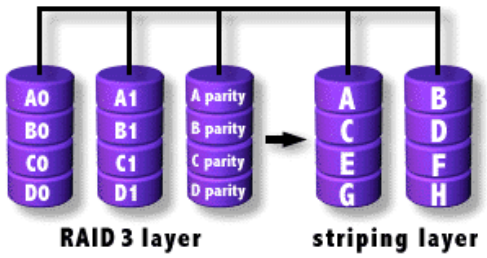
37

RAID

- RAID Level 53
  - Characteristics/Advantages
    - RAID 53 should really be called "RAID 03" because it is implemented as a striped (RAID level 0) array whose segments are RAID 3 arrays
    - RAID 53 has the same fault tolerance as RAID 3 as well as the same fault tolerance overhead
    - High data transfer rates are achieved thanks to its RAID 3 array segments
    - High I/O rates for small requests are achieved thanks to its RAID 0 striping
    - Maybe a good solution for sites who would have otherwise gone with RAID 3 but need some additional performance boost
  - Disadvantages
    - Very expensive to implement
    - All disk spindles must be synchronized, which limits the choice of drives
    - Byte striping results in poor utilization of formatted capacity

38

RAID 53: High I/O Rates and Data Transfer Performance



RAID Level 53 requires a minimum of 5 drives to implement

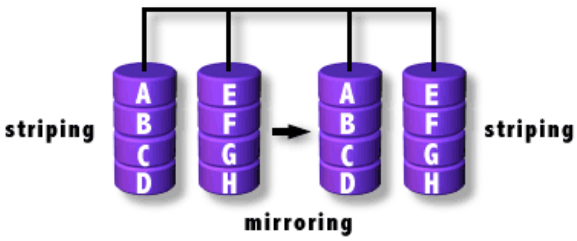
39

RAID

- RAID Level 0+1
  - Characteristics/Advantages
    - RAID 0+1 is implemented as a mirrored array whose segments are RAID 0 arrays
    - RAID 0+1 has the same fault tolerance as RAID level 5
    - RAID 0+1 has the same overhead for fault-tolerance as mirroring alone
    - High I/O rates are achieved thanks to multiple stripe segments
    - Excellent solution for sites that need high performance but are not concerned with achieving maximum reliability
  - Disadvantages
    - RAID 0+1 is NOT to be confused with RAID 10. A single drive failure will cause the whole array to become, in essence, a RAID level 0 array
    - Very expensive / High overhead
    - All drives must move in parallel to proper track lowering sustained performance
    - Very limited scalability at a very high inherent cost
  - Recommended Applications
    - Imaging applications
    - General fileserver

40

RAID 0+1: High Data Transfer Performance



RAID Level 0+1 requires a minimum of 4 drives to implement

41

Physical Placement of Data

- Logical RAIDs (Software RAID)
  - Not connected to a single controller
  - Strip the data among the disks in the networks
  - Filesystem is responsible for both distributing the data and maintaining the desired tolerance level
  - Behave like RAID5

42

Physical Placement of Data

- Stripe Groups
  - Disadvantages of one groups having a very large disks
    - Many small write operations (can't use bandwidth of the disks)
    - Node's limitation
    - Probability of a failure increases
  - Solution
    - Grouping of disks

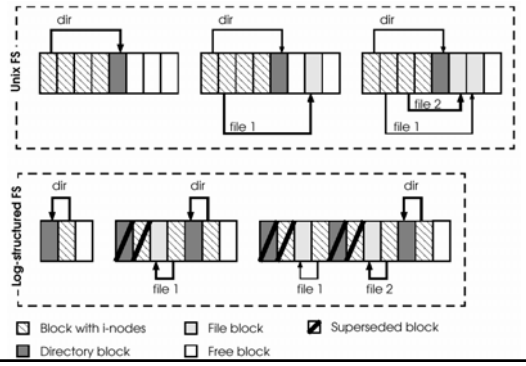
43

Physical Placement of Data

- Log-Structured Filesystem
  - Idea
    - most write operations are done sequentially
  - Reduces the small-write problem
  - Based on the assumption that caches obtain very high read
    - increase the disk performance
  - Behaves as log
  - Differences between traditional Unix FS & log-structured ones
    - all write are done sequentially in the log-structured FS
    - no such thing happens on a traditional one
    - ease the disk performance

44

Differences Between Traditional Unix FS and Log-structured Ones



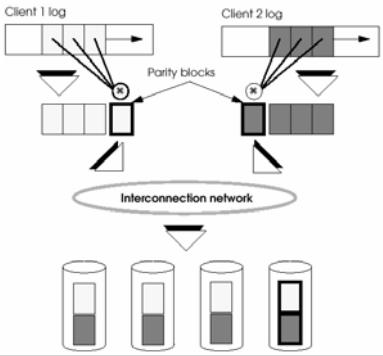
45

Physical Placement of Data

- Solving the Small-Write Problem
  - Basic Idea
    - Mixing the log-structured filesystem and the logical RAID so that small-writes never occur
  - Using the cache to avoid writing until a large block is available, logging the parity, or building a two level RAID

46

The Log-structured Filesystem as a Solution to the Small-write Problem in a Cluster of Workstations



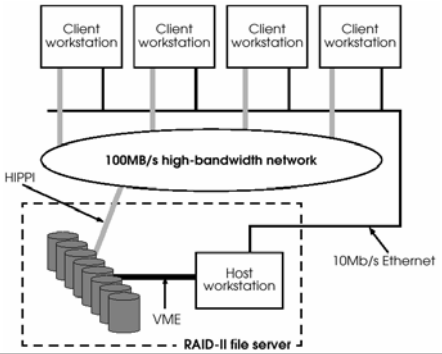
47

Physical Placement of Data

- Network-Attached Devices
  - One of the problems in a file server
    - The bandwidth of this disk is limited to the bandwidth of the memory in the server
    - The operations in the server become I/O bottleneck
  - To solve
    - Network-attached devices
      - I/O devices should be connected to a host & to a very high-bandwidth network
    - Example
      - RAID II system
      - Global File System

48

Example of a RAID-II File Server and Its Clients



49

Physical Placement of Data

- Network-Attached Devices
  - RAID-II system
    - Three components
      - A high bandwidth RAID
      - The high bandwidth network
      - A host node
  - Global File System
    - A prototype design for a distributed filesystem
    - Contribution
      - The locking mechanism

50