Section 12: File Systems, Journaling

CS 162

August 3, 2020

Contents

1 Vocabulary

- **Transaction** - A transaction is a unit of work within a database management system. Each transaction is treated as an indivisible unit which executes independently from other transactions. The ACID properties are usually used to describe reliable transactions.
- **ACID** - An acronym standing for the four key properties of a reliable transaction.
  - *Atomicity* - The transaction must either occur in its entirety, or not at all.
  - *Consistency* - Transactions must take data from one consistent state to another, and cannot compromise data integrity or leave data in an intermediate state.
  - *Isolation* - Concurent transactions should not interfere with each other; it should appear as if all transactions are serialized.
  - *Durability* - The effect of a committed transaction should persist despite crashes.
- **Idempotent** - An idempotent operation can be repeated withiout an effect after the first iteration.
- **Logging File System** - A logging file system (or journaling file system) is a file system in which all updates are performed via a transaction log ("journal") to ensure consistency in case the system crashes or loses power. Each file system transaction is written to an append-only redo log. Then, the transaction can be committed to disk. In the event of a crash, a file system recovery program can scan the journal and re-apply any transactions that may not have completed successfully. Each transaction must be idempotent, so the recovery program can safely re-apply them.

2 Inode-Based File System (7/29)

1. What are the advantages of an inode-based file system design compared to FAT?

   Fast random access to files. Support for hard links.

2. What is the difference between a hard link and a soft link?

   Hard links point to the same inode, while soft links simply list a directory entry. Hard links use reference counting. Soft links do not and may have problems with dangling references if the referenced file is moved or deleted. Soft links can span file systems, while hard links are limited to the same file system.

3. Why do we have direct blocks? Why not just have indirect blocks?

   Faster for small files.

4. Consider a file system with 2048 byte blocks and 32-bit disk and file block pointers. Each file has 12 direct pointers, a singly-indirect pointer, a doubly-indirect pointer, and a triply-indirect pointer.

   (a) How large of a disk can this file system support?

   2^{32} blocks x 2^{11} bytes/block $= 2^{43} =$ 8 Terabytes.

   (b) What is the maximum file size?

   There are 512 pointers per block (i.e. 512 4-byte pointers in 2048 byte block), so: blockSize $\times$ (numDirect + numIndirect + numDoublyIndirect + numTriplyIndirect)

$$\begin{aligned} 2048 \times (12 + 512 + 512^2 + 512^3) &= 2^{11} \times (2^2 \times 3 + 2^9 + 2^{9\times 2} + 2^{9\times 3}) \\ &= 2^{13} \times 3 + 2^{20} + 2^{29} + 2^{38} \\ &= 24K + 513M + 256G \end{aligned}$$

5. Rather than writing updated files to disk immediately, many UNIX systems use a delayed *write-behind policy* in which dirty disk blocks are flushed to disk once every x seconds. List two advantages and one disadvantage of such a scheme.

   Advantage 1: The disk scheduling algorithm (i.e. SCAN) has more dirty blocks to work with at any one time and can thus do a better job of scheduling the disk arm.
   Advantage 2: Temporary files may be written and deleted before data is written to disk.
   Disadvantage: File data may be lost if the computer crashes before data is written to disk.

6. List the set of disk blocks that must be read into memory in order to read the file `/home/cs162/test.txt` in its entirety from a UNIX BSD 4.2 file system (10 direct pointers, a singly-indirect pointer, a doubly-indirect pointer, and a triply-indirect pointer). Assume the file is 15,234 bytes long and that disk blocks are 1024 bytes long. Assume that the directories in question all fit into a single disk block each. Note that this is not always true in reality.

1. Read in file header for root (always at fixed spot on disk).
2. Read in first data block for root (/).
3. Read in file header for home.
4. Read in data block for home.
5. Read in file header for cs162.
6. Read in data block for cs162.
7. Read in file header for test.txt.
8. Read in data block for test.txt.

9. - 17. Read in second through 10th data blocks for test.txt.

18. Read in indirect block pointed to by 11th entry in test.txt's file header.

19. - 23. Read in 11th – 15th test.txt data blocks. The 15th data block is partially full.

3 Comparison of File Allocation Strategies

In lecture three file allocation strategies were discussed: (a) Indexed files. (b) Linked files. (c) Contiguous (extent-based) allocation.

Each of these strategies has advantages and disadvantages, which depend on the goals of the file system and the expected file access patterns. For each of the following situations, rank the three designs in order of best to worst. Give a reason for your ranking.

1. You have a file system where the most important criteria is the performance of sequential access to very large files.

1. c (extent-based)
2. b (linked)
3. a (indexed)

It is easy to see that (c) is the best structure for sequential access to very large files, since in (c) files are contiguously allocated and the next block to read is physically the next on the disk. No seek time to find the next block, and each block will be read sequentially as the disk head moves.

Both (b) and (a) require some look up operation in order to know where the next block is. However, (a) may be slightly more expensive, since for "very large files", multiple disk accesses are required to read the indirect blocks.

2. You have a file system where the most important criteria is the performance of random access to very large files.

1. c (extent-based)
2. a (indexed)
3. b (linked)

(c) is still the best structure here: just need to use an offset.

(a) will probably need to look at some levels of indirect blocks in order to find the right block to access (we are dealing with very large files).

(b) is absolutely the worst structure. In fact, in order to find a random block, we will need to traverse a linked list of blocks, which will take a time linear in the offset size.

3. You have a file system where the most important criteria is the utilization of the disk capacity (i.e. getting the most file bytes on the disk).

1. b (linked)
2. a (indexed)
3. c (extent-based)

(c) can suffer heavily of external fragmentation, especially for large files. So it is not the best structure for getting the most bytes on the disk, since lots of space will be wasted. However, for small files and large block size, (c) might prove to be better than (a) and (b).

(a) and (b) stuctures are generally more suitable for this question. The metadata overhead for (b) is likely to be smaller than the one for (a), since it only needs pointers to the next allocated block rather than an entire block (or blocks) which may or may not be totally used.

4 Logs and Journaling

You create two new files, F_1 and F_2, right before your laptop's battery dies. You plug in and reboot your computer, and the operating system finds the following sequence of log entries in the file system's journal.

1. Find free blocks x_1, x_2, $\dots$, x_n to store the contents of F_1, and update the free map to mark these blocks as used.
2. Allocate a new inode for the file F_1, pointing to its data blocks.
3. Add a directory entry to F_1's parent directory referring to this inode.
4. *Commit*
5. Find free blocks y_1, y_2, $\dots$, y_n to store the contents of F_2, and update the free map to mark these blocks as used.
6. Allocate a new inode for the file F_2, pointing to its data blocks.

What are the possible states of files F_1 and F_2 *on disk* at boot time?

- File F_1 may be fully intact on disk, with data blocks, an inode referring to them, and an entry in its parent directory referring to this inode.
- There may also be no trace of F_1 on disk (outside of the journal), if its creation was recorded in the journal but not yet applied.
- F_1 may also be in an intermediate state, e.g., its data blocks may have been allocated in the free map, but there may be no inode for F_1, making the data blocks unreachable.
- F_2 is a simpler case. There is no *Commit* message in the log, so we know these operations have not yet been applied to the file system.

Say the following entries are also found at the end of the log:

7. Add a directory entry to F_2's parent directory referring to F_2's inode.
8. *Commit*

How does this change the possible states of file F_2 on disk at boot time?

The situation for F_2 is now the same as F_1: the file and its metadata could be fully intact, there could be no trace of F_2 on disk, or any intermediate between these two states.

Say the log contained only entries (5) through (8) shown above. What are the possible states of file F_1 on disk at the time of the reboot?

We can now assume that F_1 is fully intact on disk. The log entries for its creation are only removed from the journal when the operation has been fully applied on disk.

What is the purpose of the *Commit* entries in the log?

- The *Commit* entry makes the creation of each file *atomic*. These changes to the file system's on-disk structures are either completely applied or not applied at all.
- The creation of a file involves multiple steps (allocating data blocks, setting up the inode, etc.) that are not inherently atomic, nor is the action of recording these actions in the journal, but we want to treat these steps as a single logical transaction.
- Appending the final *Commit* entry to the log (a single write to disk) *is* assumed to be an atomic operation and serves as the "tipping point" that guarantees the transaction is eventually applied.

When recovering from a system crash and applying the updates recorded in the journal, does the OS need to check if these updates were partially applied before the failure?

No. The operation for each log entry (e.g., updating an inode or a directory entry) is assumed to be *idempotent*. This greatly simplifies the recovery process, as it is safe to simply replay each committed transaction in the log, whether or not it was previously applied.