Operating Systems 2021, Assignment 3: File System Implementation

Deadline: Friday, May 21, 18:00

1 Introduction

A disk can be accessed as an array of disk blocks. Typical disk block sizes are 512, 1024, 2048, 4096 or 8192 bytes. In general, larger file systems employ larger disk blocks in order to keep things manageable. In order to store files and directories in such an array of blocks, we need to think about how to organize the data. In which of the free blocks will we write a given file? Which blocks on the device are actually free (not in use)? How can we find out in which blocks the file's content is located? How do we store other data about a file, such as its permissions and time of last modification? Finally, how do we store directories? Several "formats" to organize such data have been devised over the years, such as FAT, NTFS, ext2, HFS and XFS. We usually refer to these formats as *file systems*.

A file system can be split in roughly two parts: the actual data and the metadata about this data. The metadata contains a table of filenames and information about these filenames such as permissions, file size and more importantly a list of disk blocks where the contents of the file can be found. Moreover, the directory structure (tree) must be stored as well!

In this final assignment we will implement parts of a simple file system, named EdFS[1]. The initial code that you are provided with only provides a framework and does not read any data from the file system. The technical design of the file system is described in Appendix A. You will work on the implementation of support to read directories, to make/remove directories, to read files and finally to create and write files. The tasks gradually increase in complexity. For the basic tasks some guidance is included, read on for this below. There is less guidance for the more difficult parts and in order to be awarded an excellent grade we expect you to be able to work out solutions on your own.

Usually implementations of file systems are done as part of an operating system, for example as kernel module. However, for this assignment we will be using the FUSE system[2]. FUSE (Filesystem in User SpacE) allows file systems to be implemented in user space. The FUSE infrastructure will handle all necessary communication with the kernel to make this possible. To facilitate development we will not be storing the file system on an actual device, but within a file. So, in fact a file on the host computer is assuming the role of disk. We will refer to this file as an *image file*. A populated image file with file content and an empty initialized image file are available from BrightSpace and should be used to test your implementation.

2 Requirements

Although the starting point you are given is written in pure C, it is allowed to use C++. We expect the following to be achieved:

- Implement support for reading (sub)directories such that the full directory tree stored on the provided reference image can be read.
- Implement support for reading files from the file system. Your implementation must be able to read all files on the provided reference file system image. It must be able to cope with reads of arbitrary size and at arbitrary offsets. We will verify this by computing the checksums of the read files and comparing these with the reference.
- Implement support for the creation and removal of (sub)directories. The user must be able to create/remove directories using *mkdir* and *rmdir*.
- Implement support for the creation of new files. This should correctly create new inodes and directory entries.

[1] Educational File System

[2] `https://github.com/libfuse/libfuse`

- Implement support for writing to files.
  - When necessary, new disk blocks must be allocated and be registered in the inode for the corresponding file.
  - The file size must be correctly updated in the inode. Changes must be persistent after re-mounting the file system.
  - We have provided a utility called *overwrite* which overwrites a given file with a specific pattern. After using this utility, it must be possible to read back the correct contents of the file (for example using *cat*) and *md5sum* must be able to compute the correct checksum before and after re-mounting the filesystem. See also the section on Testing below.
  - The *overwrite* utility must work on files in both the root directory as well as subdirectories.
- Implement support for file truncation, so it becomes possible to fully overwrite files.
- For all changes made to the file system holds:
  - Files and directories must be stored on disk according to the specification.
  - Changes must be persistent after re-mounting the file system.
  - *fsck.edfs* must always pass (report no errors) when run on an unmounted file system, also after the file system has been modified.
- The code that you will write must be modular and avoid extensive code duplication. This will be assessed and reflected in the Quality component of your grade.
- *Important!* Make sure to use extensive error handling in your code so that your code detects various kinds of failures. See below for a list of commonly used error codes.

3 Submission and Grading

You may work in teams of at most 2 persons. Your submission should consist of the source code of the FUSE program with your modifications. *Make sure that all files that you have modified contain your names and student IDs.* Put all files to deliver in a separate directory (e.g. `lab3`) and remove any object files and binaries. *And please in particular* DO NOT *include image files in your tar file.* Finally create a gzipped tar file of this directory:

```
tar -czvf sXXXXXXX-sYYYYYYY-lab3.tar.gz lab3/
```

Substitute `XXXXXXX` and `YYYYYYY` with your student IDs. **If the tar file is larger than 40KiB (or 1 MiB if a git repository is included) you have likely included unnecessary files. Please double check.**
Submit the tar-archive through the BrightSpace submission site. Also note your names and student IDs in the text box in the submission website. *Please, ensure only one team member submits the assignment, such that there is a single submission per team!*

Deadline: Friday, May 21, 2021, 18:00.

Notes:

- For those who need the weekend to finalize things: BrightSpace submission is open until Sunday, May 23, 23:59. After this time submissions will be considered late. Note that no questions will be answered on the mailing list after Friday 18:00.
- All source code that is submitted will be subjected to automatic plagiarism checks. Cases of plagiarism will be reported to the board of examiners.
- As with all other course work, keep assignment solutions to yourself. Do not post the code on public Git or code snippet repositories where it can be found by other students.
- Test on the university Linux computers before handing in. In the case of disputes, the university Linux installation is used as reference.
- We may always invite teams to elaborate on their submission in an interview in case parts of the source code need further explanation.

4 Assessment

The maximum grade that can be obtained for this assignment is 10. The points are distributed as follows:

- **[1.5 out of 10]** Code layout and quality
- **[1.5 out of 10]** Reading directories
- **[2.0 out of 10]** Reading files
- **[2.0 out of 10]** Creating/removing directories
- **[2.0 out of 10]** Writing to (existing) files
- **[0.5 out of 10]** Creating files
- **[0.5 out of 10]** Truncating files

For code quality the following is considered: structure and modularity, consistency of the indentation and brace style, comments where these are required, quality of error handling.

5 FUSE

5.1 Using FUSE

To use FUSE on your own computer, make sure to install the `libfuse-dev` package (Debian/Ubuntu). We have tested the assignment on the university computers (Ubuntu 18.04), where FUSE version 2.9.7 is used. We expect no problems when a newer version of FUSE is used, but we have not tested this.

If you want to work remotely on the university computers, then you need to enable the following environment before working on the assignment:

```
source /vol/share/groups/liacs/scratch/os2021/os2021.bashrc
```

The starting point can be compiled using the supplied `Makefile`. An initialized file system image can be obtained from BrightSpace. To be able to mount this file system, first a mountpoint has to be created. **Important: on the university computers this mountpoint *cannot* be located on a network share, such as your home directory**. What does work is creating a directory under */tmp*, for example */tmp/testmyusername*. Now the filesystem can be mounted by running:

```
./edfuse myimage.img /tmp/testmyusername
```

Enter the mountpoint to browse the contents of the filesystem. To unmount the filesystem use:

```
fusermount -u /tmp/testmyusername
```

To facilitate debugging it helps to provide the `-f` and `-s` options to *edfuse*:

```
./edfuse -f -s myimage.img /tmp/testmyusername
```

This way, you can also run *edfuse* from within a debugger, such as gdb, to debug your code.

5.2 FUSE API

The FUSE API is structured around the concept of a Virtual File System (VFS). Using a VFS, we have an overview of the entire file system of a computer system (which comprises multiple file systems) as well as a generic interface to the functions of the correct file system implementation in order to carry out the desired operation on a file or directory. The interface makes it easy to implement file systems which is essentially done by implementing the functions in the FUSE file operations structure.

In the starting point that is provided to you, you can find the FUSE file operations structure at the bottom of the `edfuse.c` file. As you can see, functions are registered here for the different functionalities of the file system, such as reading directories, writing files and creating directories. This file operations structure is passed to the FUSE library during initialization in the main function.

The function prototypes for the different operations can be found in the FUSE header file, but these that are needed have already been stubbed out for you in the `edfuse.c` source code file. The arguments required for the different operations are straightforward. In most cases the file or directory to be operated on is specified using a `path` string.

5.3 Assignment Structure

Note that next to `edfuse.c` the starting point contains:

- `edfs.h` general EdFS definitions, in particular struct definitions for the different data structures.
- `edfs-common.[ch]` generic EdFS routines that would be shared with other EdFS utilities.

We recommend that you adhere to this structure. When writing a new function consider whether this function is specific to the FUSE implementation or is more generic. In case of the latter, add it to `edfs-common.[ch]`.

Several functions to help you implement the file system code are already provided, such as functions to read and write inodes, to manipulate the inode table, `edfs_get_parent_inode`, FUSE functions for the `getattr` and `open` calls. The FUSE function for the `readdir` call has been partly implemented and the same holds for `edfs_find_inode`. The implementation of these two functions must be completed by you. Also, do not take these functions for granted. Make sure to study these functions and understand how they can be used.

5.4 Resources

The following resources may be of use when getting up to speed with the FUSE API:

- FUSE API documentation: `http://libfuse.github.io/doxygen/`
- Documentation about the different FUSE operations: `http://libfuse.github.io/doxygen/structfuse__operations.html`
- FUSE tutorial `http://www.cs.nmsu.edu/~pfeiffer/fuse-tutorial/`

6 Error Codes

When writing the file system code we ask you to make extensive use of error handling and to return appropriate error codes when errors are detected. Common error codes can be found in `errno.h`, use `man errno`. To help you, the following error codes are the most common:

- `ENOSPC`, no space left on device (file system is full).
- `EINVAL`, invalid value / argument specified, for instance when a given filename is invalid (too long or contains invalid characters).
- `EIO`, I/O error.
- `ENOENT`, entry does not exist.
- `ENOTDIR`, entry is not a directory.
- `EISDIR`, entry is a directory (and not a regular file).
- `ENOTEMPTY`, the directory is not empty.
- `ENOSYS`, the function is not implemented.
- `EEXIST`, entry already exists.
- `EFBIG`, file too large.

Return values that signal failure conditions are typically negative. The above error codes are positive numbers, so usually the negative number is returned: `-EEXIST`.

7 Testing

In this section, we give some guidance on how to test your implementation. In order to do so, we have provided several files and utilities. Note that after modifying the file system (creating/removing directories, writing files), it is important to also unmount the file system and mount it again to verify the made changes are persistent. To verify the file system is not corrupted after making changes to it, you must use the *fsck.edfs* utility. This utility ensures that the data on disk (so actually within the file system image) is correct and not corrupted. It should report no errors. Though, remember that the file system check might not catch every possible error!

The binary executable file for *fsck.edfs* can be obtained from BrightSpace or from `/vol/share/groups/liacs/scratch/os2021/prefix/bin`. If your computer runs a Linux distribution from the last 3 years, the binary will most probably work. If not, contact us on the mailing list and we will look into compiling a different binary. We cannot supply the source code. (In case you do not trust binaries from us, run the binaries on a university computer instead or use a virtual machine).

7.1 Reading directories

You will have to start with the implementation of support for reading directories. To test your implementation, use the populated image file provided on BrightSpace. You should be able to inspect a hierarchy of subdirectories using *ls*, *cd*, *find* and so on. A textual representation of the hierarchy as stored in the EdFS image is also available on BrightSpace.

7.2 Creating and Removing Directories

To test the support for creating and removing (sub)directories, you can use the *mkdir* and *rmdir* utilities. These take the directory to create/remove as an argument. You should test creating subdirectories in the root directory of the EdFS file system as well as creating subdirectories in subdirectories (nesting).

7.3 Reading Files

Testing support for reading files can be done using *cat*, but also by computing the MD5 checksum using the *md5sum* command. The following table contains correct MD5 checksums and file sizes for the files present on the populated file system image:

small.txt	`a4195c1cd88942fdcbc747029db5fb5a`	18 bytes
large.txt	`acbec5cf4d59cecdf22ef5129beec07f`	21959 bytes
file1.txt	`6b8d91211b247b4eef395dc8789ed52d`	71 bytes
file2.txt	`7320de03916e08e86693ad75ef8800dc`	1420 bytes
file3.txt	`9fdb85c793cf43095c11acff998b9a2f`	16330 bytes
file4.txt	`9fdb85c793cf43095c11acff998b9a2f`	16330 bytes
file5.txt	`93b47be308f91225e4716ac09f839609`	1065 bytes

The checksums for the `1.txt` to `16.txt` files can be found in the textual representation of the populated file system image (see BrightSpace).

Note that with these utilities your FUSE module will always get read requests for blocks of 4 KiB. This is due to the kernel's page cache. To be able to test reads of different sizes, you need to circumvent the kernel page cache. This can be achieved using direct I/O. For example, the following command tests reads in blocks of 256 bytes and outputs the checksum:

```
dd if=file1.txt bs=256 iflag=direct | md5sum
```

7.4 Writing Files

For testing file write support a custom *overwrite* utility is provided, which can be found at the same locations as *fsck.edfs*. You need to specify an *existing* file name as an argument. *Be careful: the utility will overwrite the specified file without asking for confirmation!* The utility will overwrite the specified file with a particular pattern. By default this pattern is 3550 bytes in length (so files that were originally larger than 3550 bytes are only partially overwritten!).

Note that the initialized file system image contains five files named from `file1.txt` to `file5.txt`. These files also contain a specific pattern. When you specify one of these files to *overwrite* (*overwrite* compares the filename), a special action will be performed that is defined for that specific file. Ensure to test your code by using *overwrite* on each of the five test files. After overwriting the file, inspect the new file size and its contents. Also compute the MD5 checksum using the *md5sum* command. The following table specifies the correct checksums (MD5) and size for the ***overwritten*** files:

file1.txt	`f4c5d24f23c0fe539cb09d526b5de899`	1278 bytes
file2.txt	`f6cc4fee814b1b0f035116d33b177ed8`	43310 bytes
file3.txt	`7adcd2bf8b139fe71de82246ff7efcf8`	16330 bytes
file4.txt	`ac66cd60692e44e0f0b5ad9cfeb7db0a`	16330 bytes
file5.txt	`bdebaadb35691a06d35b86bd258ddb54`	15265 bytes

7.5 Creating and Truncating Files

The creation of new files can be tested using the default *echo* and *dd* commands. The *touch* command can be useful to test solely file creation (ignore the FUSE warning on not being able to set times). Finally, file truncation can be tested with, for instance, the regular *dd* and *truncate* command line utilities.

8 Getting Started

To help you get started, we will be giving some guidance on the different subtasks in this section. Read the specification of the file system in Appendix A together with this guidance to fully understand what needs to be done before starting implementation.

8.1 Reading Directories

An inode either describes a file or a directory. In the case of a directory, at most both direct blocks may be allocated. Within these blocks the directory entries are stored. To read directories the implementation of the function `edfuse_readdir` must be completed: at the TODO marker you need to write code that reads the allocated disk blocks and extracts directory entries. For every valid directory entry the filler function must be called.

Additionally, you need to complete the implementation of the `edfs_find_inode` function. At the TODO marker you need to write code to, again, visit all valid directories. So, make sure to write a generic function which visits directories of a given inode! You can then reuse this later (you will need such a function more often).

8.2 Reading Files

Completing the `edfs_find_inode` function, as described in the previous subsection, is a prerequisite for working on the reading of files. In fact, this is the first function you will need to call from `edfuse_read`. You need to find the inode in order to determine whether the given path exists and you need the inode to obtain the block numbers of the data blocks where the file's data is stored.

The following arguments are provided to the read function: `path` which is the path to perform the read operation from, `buf` which is the buffer in which the read data should be stored, `size` is the number of bytes to read, `offset` specifies the position in the file where the bytes should be read and finally `fi` provides some information about the file (which does not need to be used).

Write a generic function which is able to translate the given file offset to a block number and an offset within this block. First think about how this translation would work (use pen and paper). Once you have a good idea, write the necessary implementation. Make sure to propose appropriate error codes if the offset is out of bounds.

It is important to take block boundaries into account. For example, consider a block size of 512 bytes. We are requested to read 100 bytes, starting at offset 500 of the block. Now, you need to read only 12 bytes from this first block and read the remainder from *another* block. Your implementation of the read function must be able to cope with reads that involve multiple disk blocks.

8.3 Creating and Removing Directories

To add the ability to create and remove (sub)directories, the functions `edfs_mkdir` and `edfs_rmdir` should be implemented. Stub functions have already been created for you and registered in the `fuse_operations` structure.

The *mkdir* function takes two arguments: a path and a mode. Because we do not support file permissions, the mode is ignored. The path indicates the full path to the new directory to be created. This path must be split into the name of the new subdirectory and its parent path (that is, the directory in which the new subdirectory should be created). Make sure to perform all necessary validations, such as whether the given name for the new directory is a valid name within EdFS and whether the parent directory exists and is a directory. Then you can create a new inode for the new directory and register this new inode in the parent directory. For the latter, you need to write a function that creates a new directory entry in the parent directory in which the name of the new directory is stored together with the new inode number. Also make this function for registering directory entries generic, you will need it more often. If the registration of the directory entry succeeds, don't forget to commit the inode by writing it to disk.

To complete this task you can use various utility functions that are already provided: `edfs_get_parent_inode`, `edfs_get_basename`, `edfs_new_inode`, `edfs_write_inode`.

For the new inode you do not have to allocate disk blocks, this should be done on demand. However, this does mean that when registering new entries in the parent directory, you may need to allocate a new disk block for the parent directory in case the current disk block is full or no disk block has yet been allocated. To allocate a block, scan the bitmap for a free (unallocated) block. Determine the block number, mark this block as allocated in the bitmap and finally register this block number in the inode.

Don't forget to write the inode to disk. Try to make these functions generic, also these functions will be re-used.

Implementing the *rmdir* function is easier. A single path is given as argument, which is the directory to remove. Look for the correct inode. Verify that the directory to remove is indeed empty. Remove the directory entry from its parent and release all other resources that were allocated for this directory. Finally, don't forget to mark the inode as free in the inode table.

8.4 Creating and Writing Files

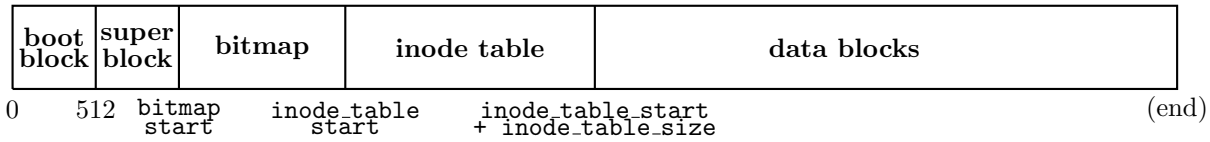
To be able to create new files, the `edfs_create` operation must be implemented. This function must create a new inode and register this inode in the parent directory (at the same time you can verify this file does not yet exist). The file size should be set to zero initially and no blocks should be allocated.

For writing to existing files the `edfs_write` operation is used. This operation is in general similar to the read operation, with two exceptions. Firstly, when necessary new blocks must be allocated. Secondly, if you have modified the file size, the inode must be updated and written to disk.

Finally, to be able fully overwrite files from the start, you must also implement the `edfs_truncate` operation. This function must set the file size to the given offset. If this offset is smaller than the current file size, redundant blocks must be released (a common operation is a truncation to zero size, which means all allocated blocks are released). The given offset can also be larger than the current file size, in which case new blocks must be allocated.

A EdFS Specification

EdFS is based on the general concept of an inode-based file system. It is comparable to, for instance, *ext2*, but has been greatly simplified. Many details will not be considered such as permissions and timestamps, and the maximum file size that will be supported is limited. We do support file systems of different sizes and using different block sizes.
The following is an overview of the file system structure:



A.1 Super block

As can be seen in this overview, the super block is always located at an offset of 512 bytes. The super block has the following structure:

```
typedef uint16_t edfs_block_t;
typedef uint32_t edfs_inumber_t;

typedef struct
{
  uint64_t magic;
  uint16_t version;

  uint16_t block_size;
  edfs_block_t n_blocks;

  uint32_t bitmap_start; /* offset from start of device; in bytes */
  uint32_t bitmap_size;  /* in bytes */

  uint32_t inode_table_start; /* offset from start of device; in bytes */
  uint32_t inode_table_size;  /* in bytes */
  uint32_t inode_table_n_inodes;

  /* Inode hosting the root directory of the file system. */
```

```
  edfs_inumber_t root_inumber;
} __attribute__((__packed__)) edfs_super_block_t;
```

The magic number must be set to the value `0x00133700f00d0034`. The version field is ignored for now. The other fields indicate the block size used in the file system (for now 512, 1024, 2048, 4096 and 8192 bytes are supported), the total number of blocks that the file system contains (this includes the blocks on which boot block, super block, bitmap and inode table are stored), start and size of the bitmap (in bytes), start and size of the inode table (in bytes). Note that the sizes of the bitmap and inode table are always rounded up to the nearest block boundary. Therefore, you must use the `n_blocks` and `inode_table_n_inodes` fields to find out the number of valid entries contained in the bitmap and inode table respectively. Finally, the last item in the superblock is the number of the inode that describes the root directory.

Note that the starting point already reads the super block for you. You do not have to do this yourself. However, in order to implement the different functionalities of the file system you must be familiar with the fields in the super block.

A.2 Bitmap

In the bitmap the state of each block that is part of the file system is maintained. A block is either free (`0`) or allocated (`1`). The `bitmap_start` field of the super block indicates the byte offset within the file system where the bitmap starts. To read the status of a given block number, the corresponding bit number must be computed. The bitmap is stored byte-wise. Within a byte, the least significant bit stores the status of block zero within that byte.

Note that the bitmap covers all blocks in the file system, also the blocks where the boot block, super block, bitmap and inode table are stored. So, the boot block is block zero and is always marked as occupied. This also holds for the blocks where the bitmap and inode table are stored. As a consequence, the offset of a block within the file system can simply be computed by multiplying the block number by the block size.

Furthermore, because block zero is always marked as occupied, the number zero is used as special identifier for unused/invalid blocks in inodes (`EDFS_BLOCK_INVALID`).

A.3 Inode table

All inodes are stored on disk consecutively starting at `inode_table_start`. The super block field `inode_table_n_inodes` indicates the number of inodes in the inode table. An inode on disk is stored according to the following structure:

```
typedef enum
{
  EDFS_INODE_TYPE_FREE = 0,
  EDFS_INODE_TYPE_FILE,
  EDFS_INODE_TYPE_DIRECTORY,
} edfs_inode_type_t;

#define EDFS_BLOCK_INVALID 0
#define EDFS_INODE_N_DIRECT_BLOCKS 2   /* NB: Increasing this value will break
                                        * compatibility.
                                        */

/* Padded to be 16 bytes in size, with 5 reserved bytes available for
 * future expansion.
 */
typedef struct
{
  edfs_inode_type_t type : 8;
  uint8_t reserved[3];

  uint32_t size;  /* file size in bytes */

  edfs_block_t direct[EDFS_INODE_N_DIRECT_BLOCKS];
```

```
  edfs_block_t indirect;
  uint16_t reserved2;
} __attribute__((__packed__)) edfs_disk_inode_t;
```

For a valid (allocated) inode the `type` field *must* be set to either the file or directory type. `direct[0]` and `direct[1]` contain pointers (in the form of disk block numbers) to the first two data blocks of the file. When unallocated, they must have the value `0`. If the file size grows beyond these two disk blocks, a single indirect block is allocated and its block number is stored in `indirect`. Block numbers to further data blocks will be stored in this indirect block. The block numbers are stored consecutively and therefore the contents of an indirect block can be seen as a small array of block numbers. Again, the special value `0` is used to indicate unallocated blocks. Note that a file may *not* contain unallocated holes.

To simplify implementation, directories *only* use the direct blocks and the indirect block is always left unallocated. The inode with number `0` is left unused and must always be marked as free. Number `1` is the first valid inode. As a consequence, we can use `0` to signify invalid inodes in directory entries.

A.4 Directories

Within the file system directories are stored by writing directory entries to allocated disk blocks. Each directory entry has the following format:

```
#define EDFS_FILENAME_SIZE (64 - sizeof(edfs_inumber_t))

typedef struct
{
  edfs_inumber_t inumber;
  char filename[EDFS_FILENAME_SIZE];
} __attribute__((__packed__)) edfs_dir_entry_t;
```

For the current version of the file system the size of a directory entry is 64 bytes. Note that a directory inode may only allocate two direct blocks to store directory entries. For a block size of 512 bytes this means that a directory can store a maximum of 16 directory entries.

The `inumber` is an inode number which is used to read the corresponding inode from the 0-indexed inode table. The `filename` is restricted to 59 bytes (excluding null-terminator) and may only contain: A-Z, a-z, 0-9, spaces (“ ”) and dots (“.”). Make sure to verify this when entering new files to the file system. A null-terminator must be stored, note that the structure allows for 60 bytes to be stored. If a directory entry is not in use, the `inumber` field must be set to zero (the invalid inode).

As an example, the root inode number indicated in the super block (often with a value of `1`) points to a valid inode with the directory type. This inode either has 0, 1 or 2 allocated blocks. Each of these allocated blocks consists of a sequence of consecutive directory entries. If the `inumber` of a directory entry is nonzero, it signifies a valid directory entry and the filename can be read from the `filename` field.