Design and Implementation of file system in BBFS to enable deduplication

Used SHA1 to hash the block to be written to disk, and stored the hash in file 'hashstore', and the corresponding data blocks in file 'blockstore'. Whenever a file is to be written using 'bb_write' function, first calculated the hash of the block (each file dived in the block size of 4KB). Then checked if the corresponding hash value is already stored in the file 'hashstore', if yes then write to the file corresponding pointer of that block. If the corresponding hash is not stored, it means that the block is new one, then make a new entry in the 'hashstore' file and the corresponding entry of data in the 'blockstore' file. The file to which the is to be written, write the pointer of the data block instead the actual data.

When reading data back from the file, 'bb_read' function look through the pointer of data block in the file and give the corresponding data of file from the 'blockstore' function.

Function Changed:

1. bb_write (in file bbfs.c)
2. bb_read (in file bbfs.c)
3. main (in file bbfs.c)
4. find_hash (newly created in bbfs.c)
5. int2str (newly created in bbfs.c)
6. str2int (newly created in bbfs.c)

find_hash:

Input: block of 4KB size

Output: if hash exists already for this block, return its pointer

Else, return -1

int2str:

Convert integer to string

str2int:

Convert string to integer

main:

Open the file 'hashstore' and 'blockstore', which contains the already hashed value and the corresponding block data

bb_write:

Modified to enable deduplication of block

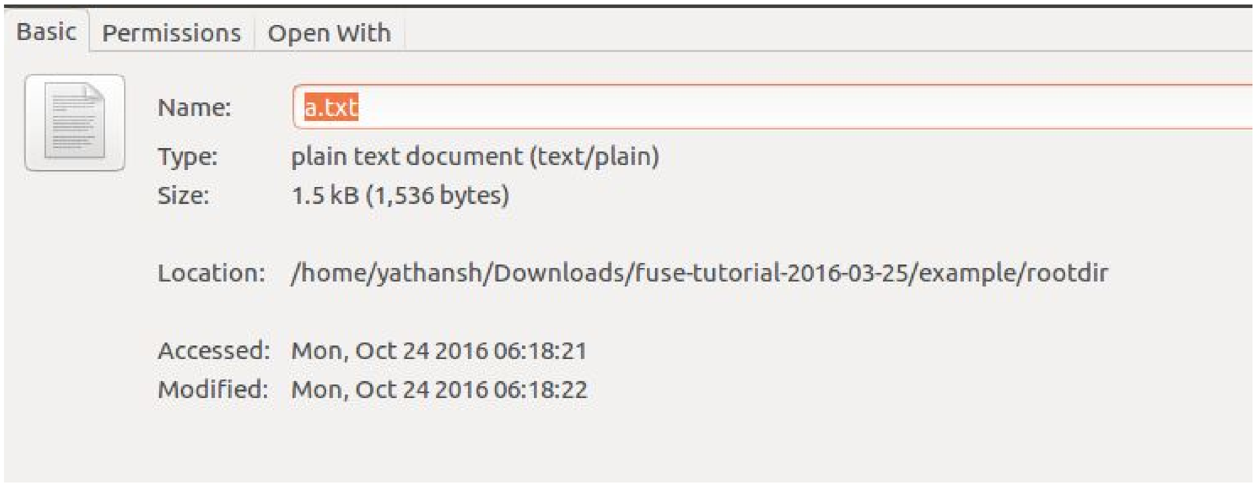
bb_read:

Modified to read block from the file which contains pointer of the 4KB block

Testing Correctness:

Case 1:
We created a single file of 20,97,152 bytes (512 blocks) with lots of repetition i.e. there are a lot of same blocks of 4KB in the file (only 3 different blocks are present)
The file size of the file in the mount directory was 20,97,152 bytes itself but that in root directory was 1536B (because the files true contents are just the pointers to its blocks)
The file where the blocks were stores (/tmp/blockstore) was of 12288 Bytes (3 *4096) only thus confirming that the only 3 distinct blocks were stored on the disk

Case 2:
We created another file of 20,97,152 bytes (512 blocks) with some blocks same as that of the above file and some new blocks (This file too had only 3 distinct blocks out of which one was common to the 1st file)
On doing this the size of both file in the mount folder was 20,97,152 bytes whereas in the root directory was 1536 B (as only the pointers are stored in the disk image of the file)
Also the size of /tmp/blockstore changed to 20480 Bytes (5*4096) denoting that 5 are stored in it (we have a total of 5 distinct blocks only, 3 in a.txt, 3 in b.txt and 1 among them common

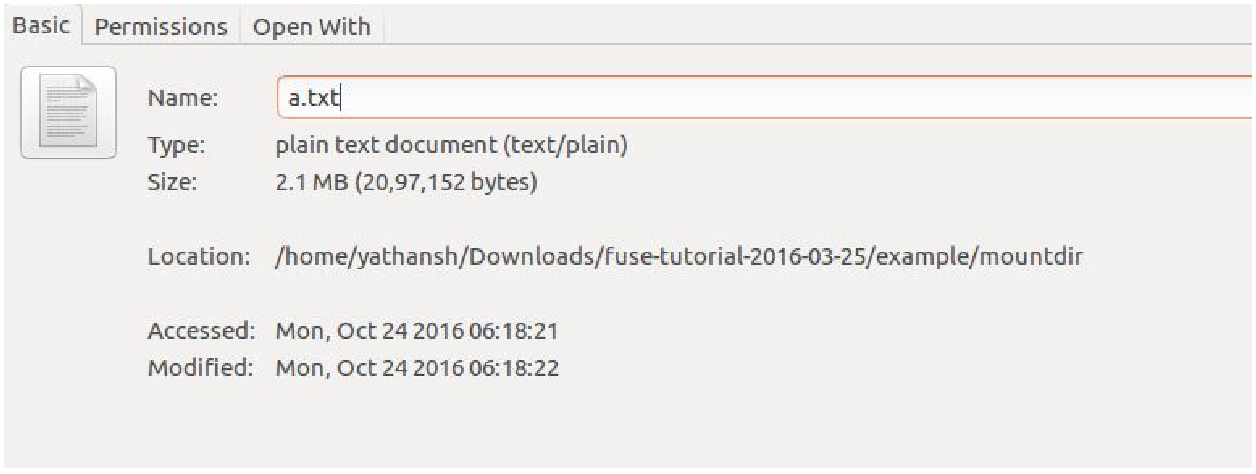


Name: a.txt
Type: plain text document (text/plain)
Size: 1.5 kB (1,536 bytes)

Location: /home/yathansh/Downloads/fuse-tutorial-2016-03-25/example/rootdir

Accessed: Mon, Oct 24 2016 06:18:21
Modified: Mon, Oct 24 2016 06:18:22

Root size of a.txt

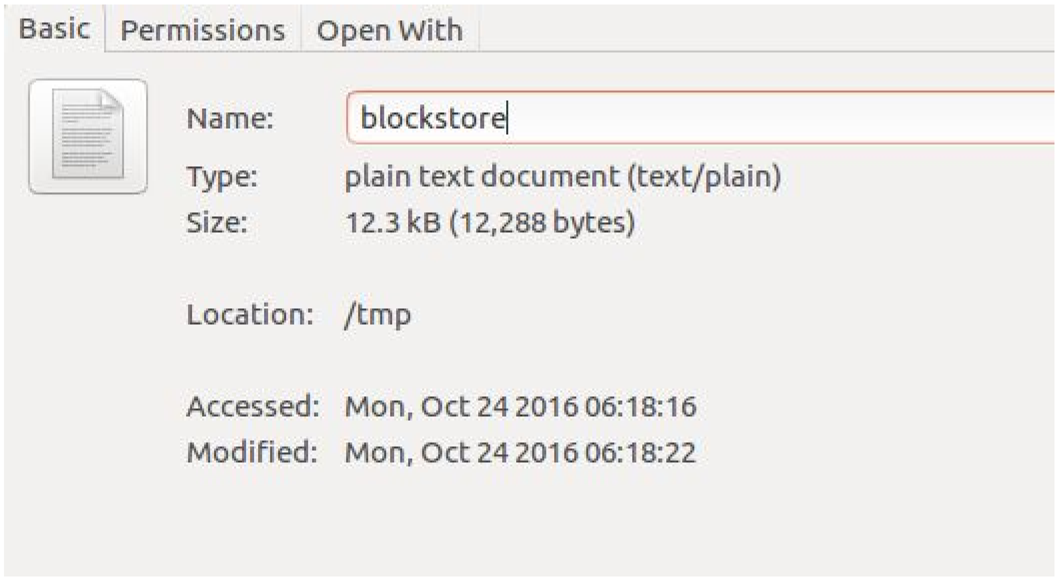


Name: a.txt
Type: plain text document (text/plain)
Size: 2.1 MB (20,97,152 bytes)

Location: /home/yathansh/Downloads/fuse-tutorial-2016-03-25/example/mountdir

Accessed: Mon, Oct 24 2016 06:18:21
Modified: Mon, Oct 24 2016 06:18:22

Mount size of a.txt

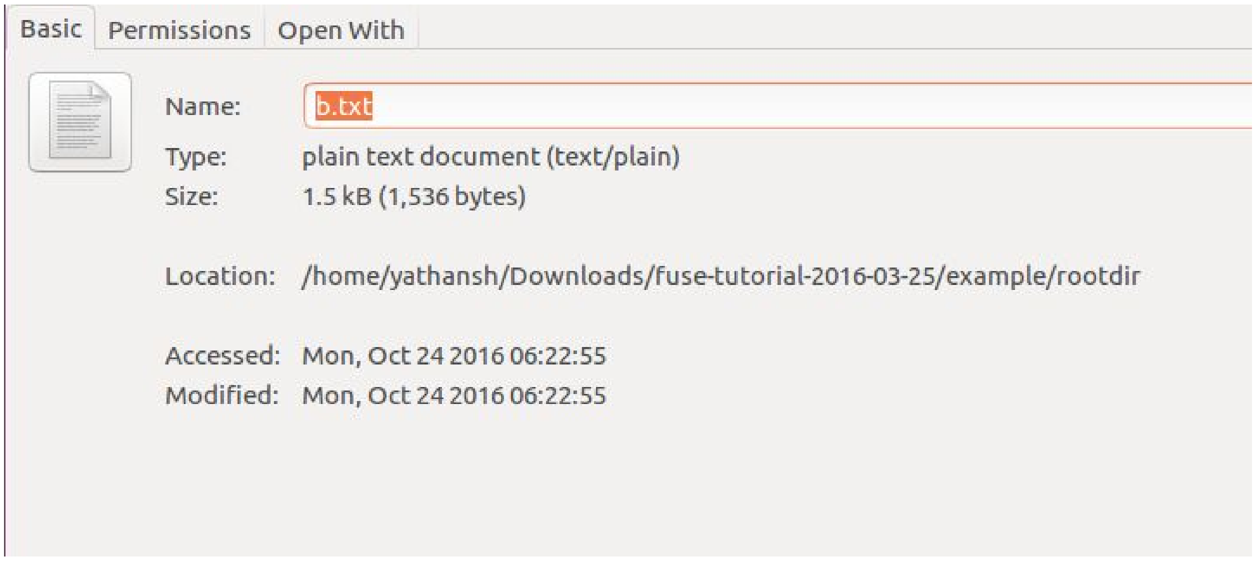


Name: blockstore
Type: plain text document (text/plain)
Size: 12.3 kB (12,288 bytes)

Location: /tmp

Accessed: Mon, Oct 24 2016 06:18:16
Modified: Mon, Oct 24 2016 06:18:22

Size of blockstore when only a.txt is stored

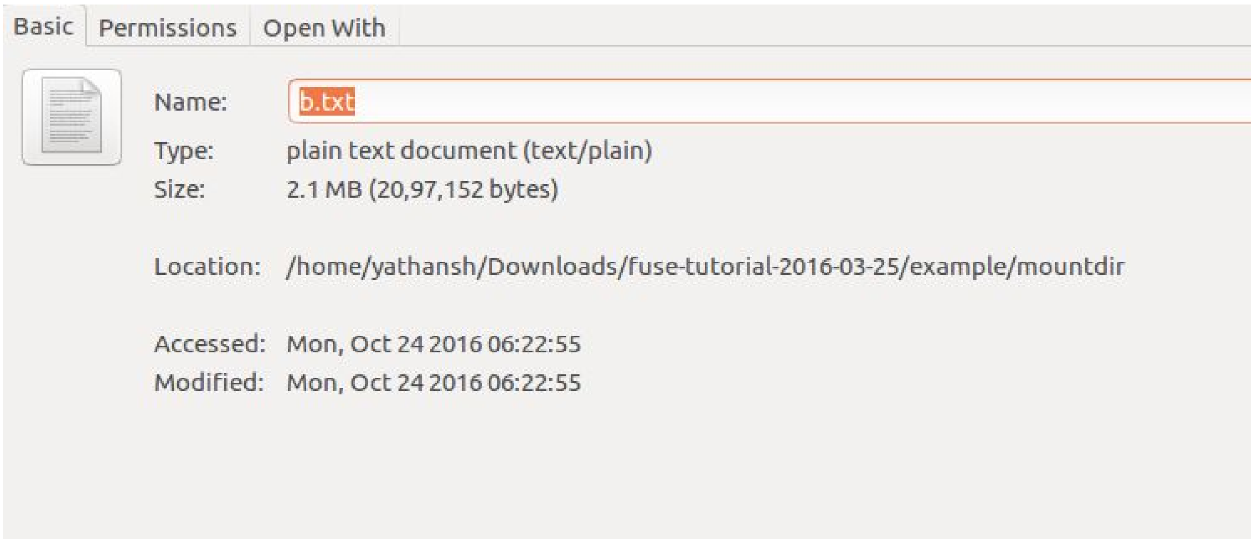

Basic | Permissions | Open With

Name: b.txt
Type: plain text document (text/plain)
Size: 1.5 kB (1,536 bytes)

Location: /home/yathansh/Downloads/fuse-tutorial-2016-03-25/example/rootdir

Accessed: Mon, Oct 24 2016 06:22:55
Modified: Mon, Oct 24 2016 06:22:55

Root size of b.txt

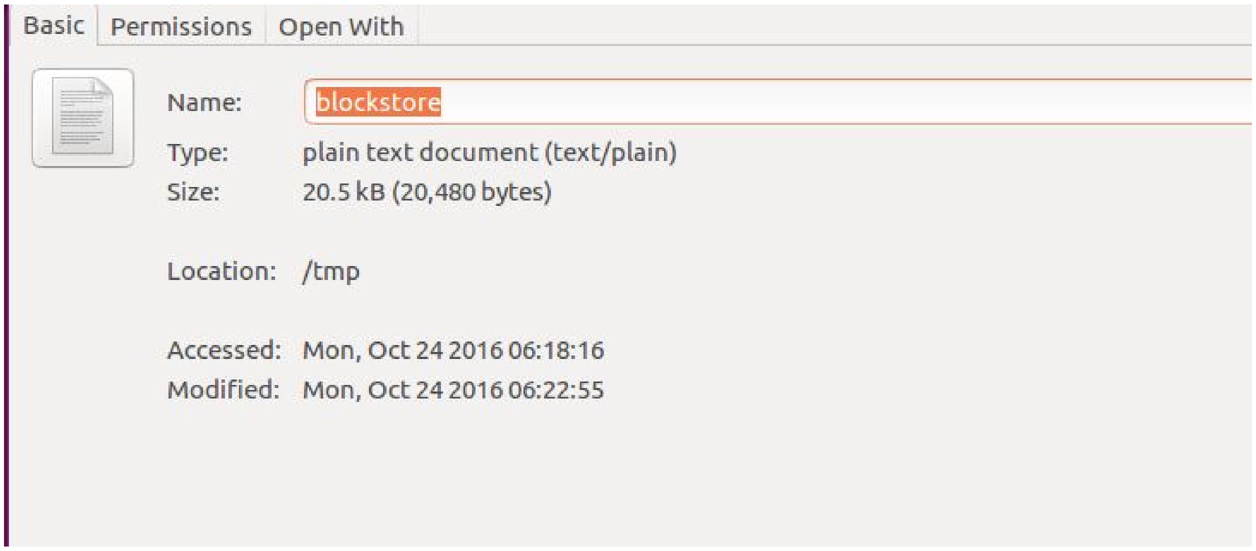

Basic | Permissions | Open With

Name: b.txt
Type: plain text document (text/plain)
Size: 2.1 MB (20,97,152 bytes)

Location: /home/yathansh/Downloads/fuse-tutorial-2016-03-25/example/mountdir

Accessed: Mon, Oct 24 2016 06:22:55
Modified: Mon, Oct 24 2016 06:22:55

Mount size of b.txt



Name: blockstore

Type: plain text document (text/plain)

Size: 20.5 kB (20,480 bytes)

Location: /tmp

Accessed: Mon, Oct 24 2016 06:18:16

Modified: Mon, Oct 24 2016 06:22:55

Size of blockstore when both a.txt and b.txt are stored