

INFO 0940

Project 4

## Assignment 4: An OS support for faster cmp

- Please extend the ext4 file system so that it can maintain the hash tree for every file
- Please implement a cmp-like application which compares the hash trees to find out diffs of two files
  - whose output is similar to “cmp -l FILE1 FILE2”
- Specification
  - Please store the hash trees on the disk
  - Please update the hash trees in kernel
    - Please do not modify the hash tree from user-space

```
$ man cmp
```

```
CMP(1)
```

```
NAME
```

```
cmp - compare two files byte by byte
```

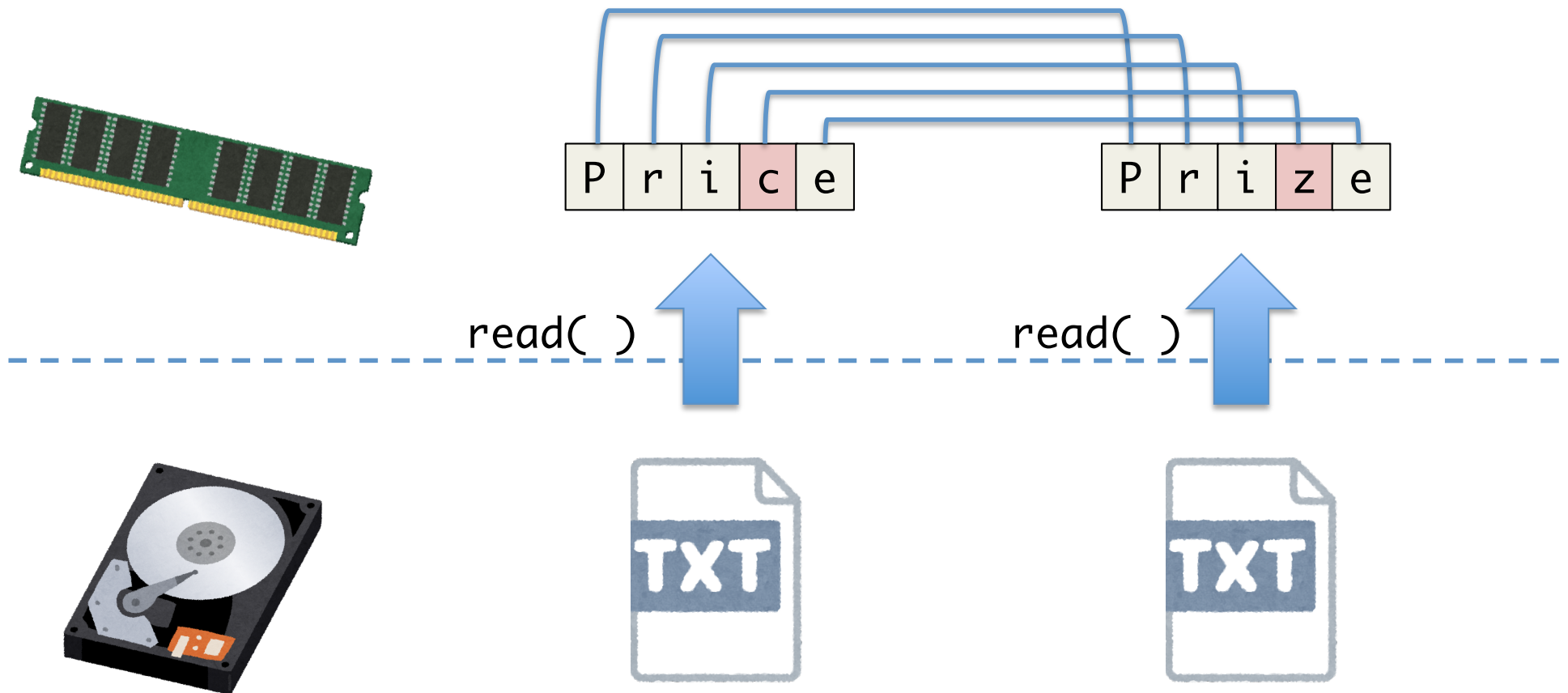
```
SYNOPSIS
```

```
cmp [OPTION]... FILE1 [FILE2 [SKIP1 [SKIP2]]]
```

```
DESCRIPTION
```

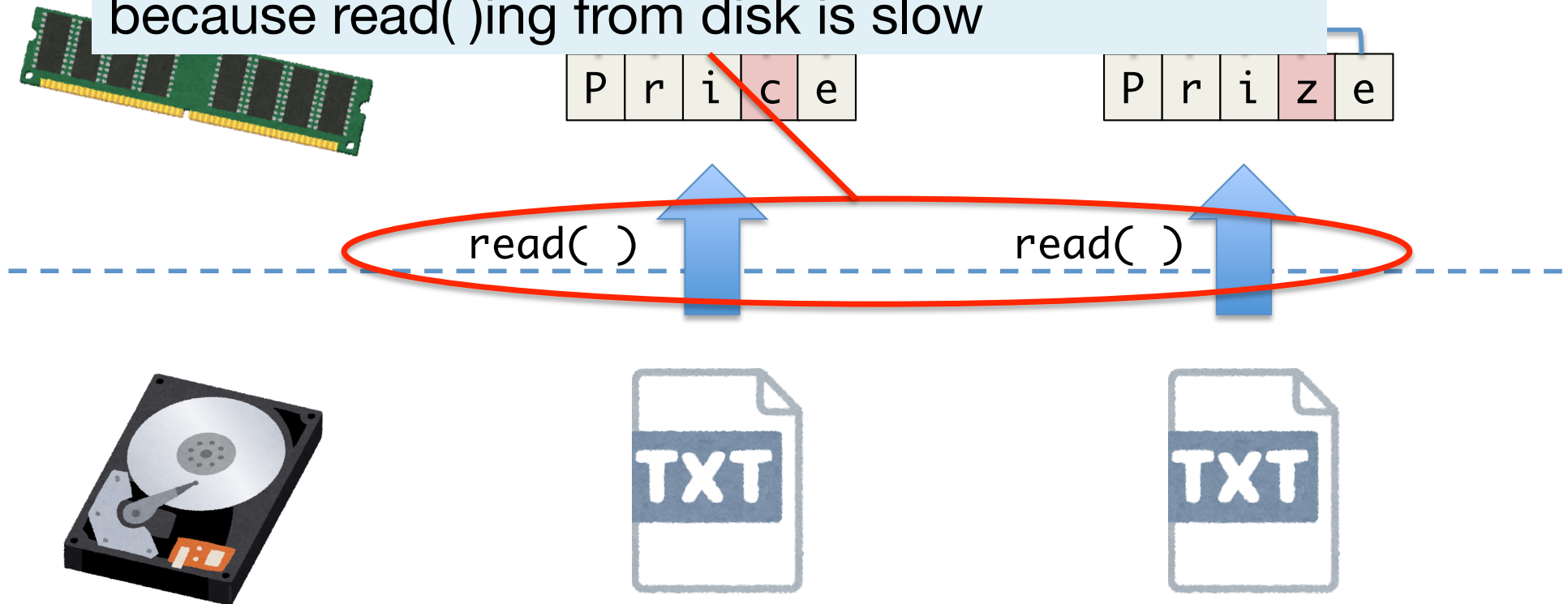
```
Compare two files byte by byte.
```

# Comparing two files



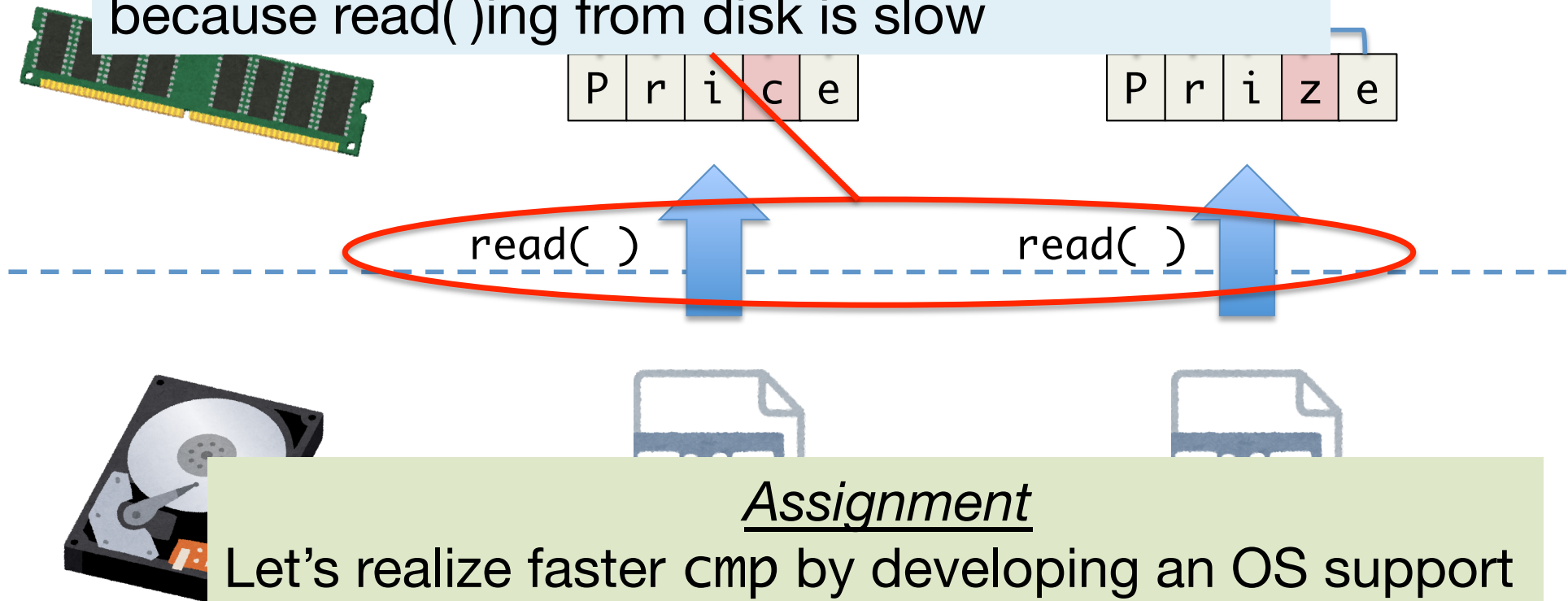
# Comparing two files

Problem : when files are big, it takes a lot of time because read( )ing from disk is slow



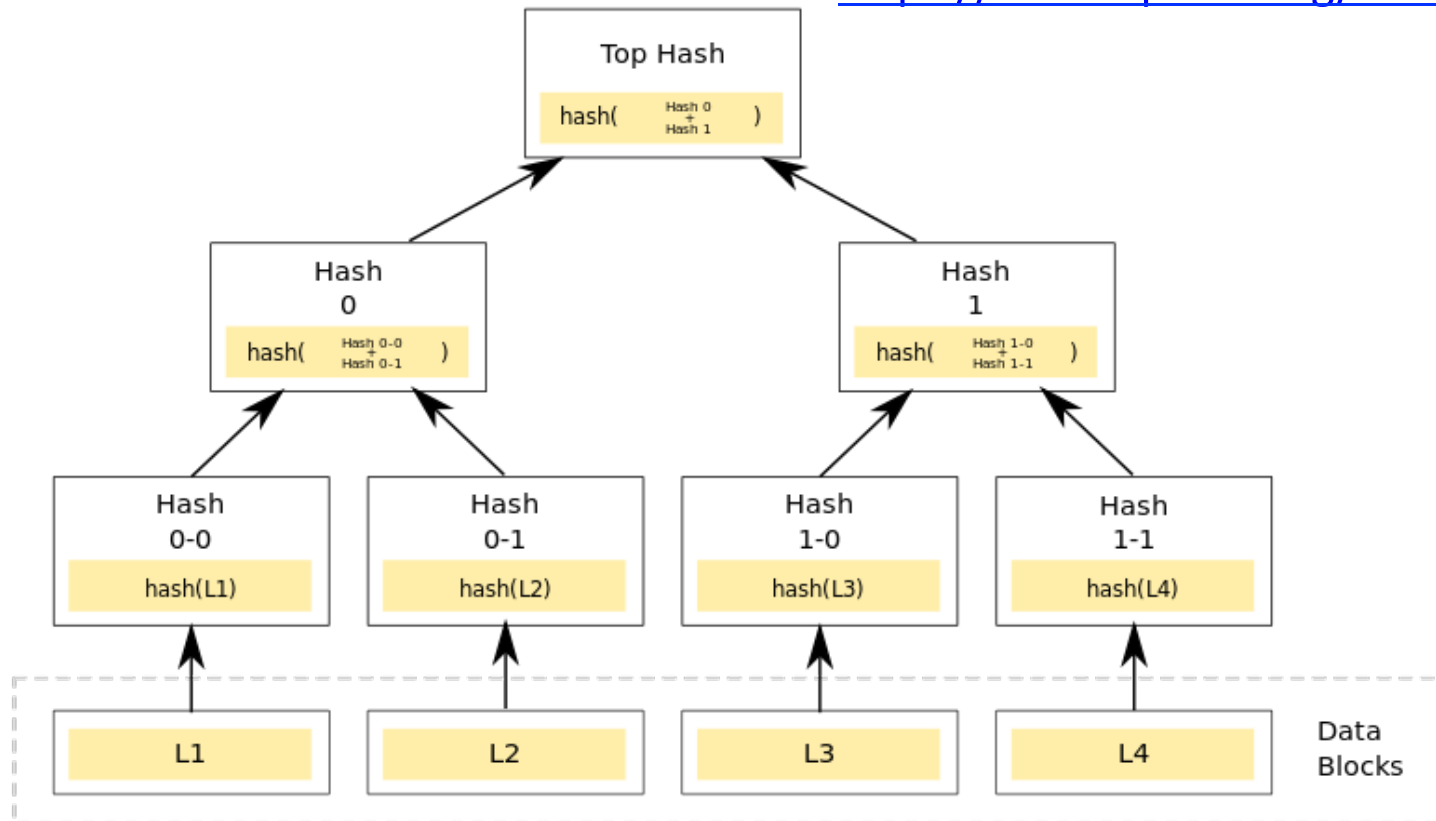
# Comparing two files

Problem : when files are big, it takes a lot of time because read( )ing from disk is slow

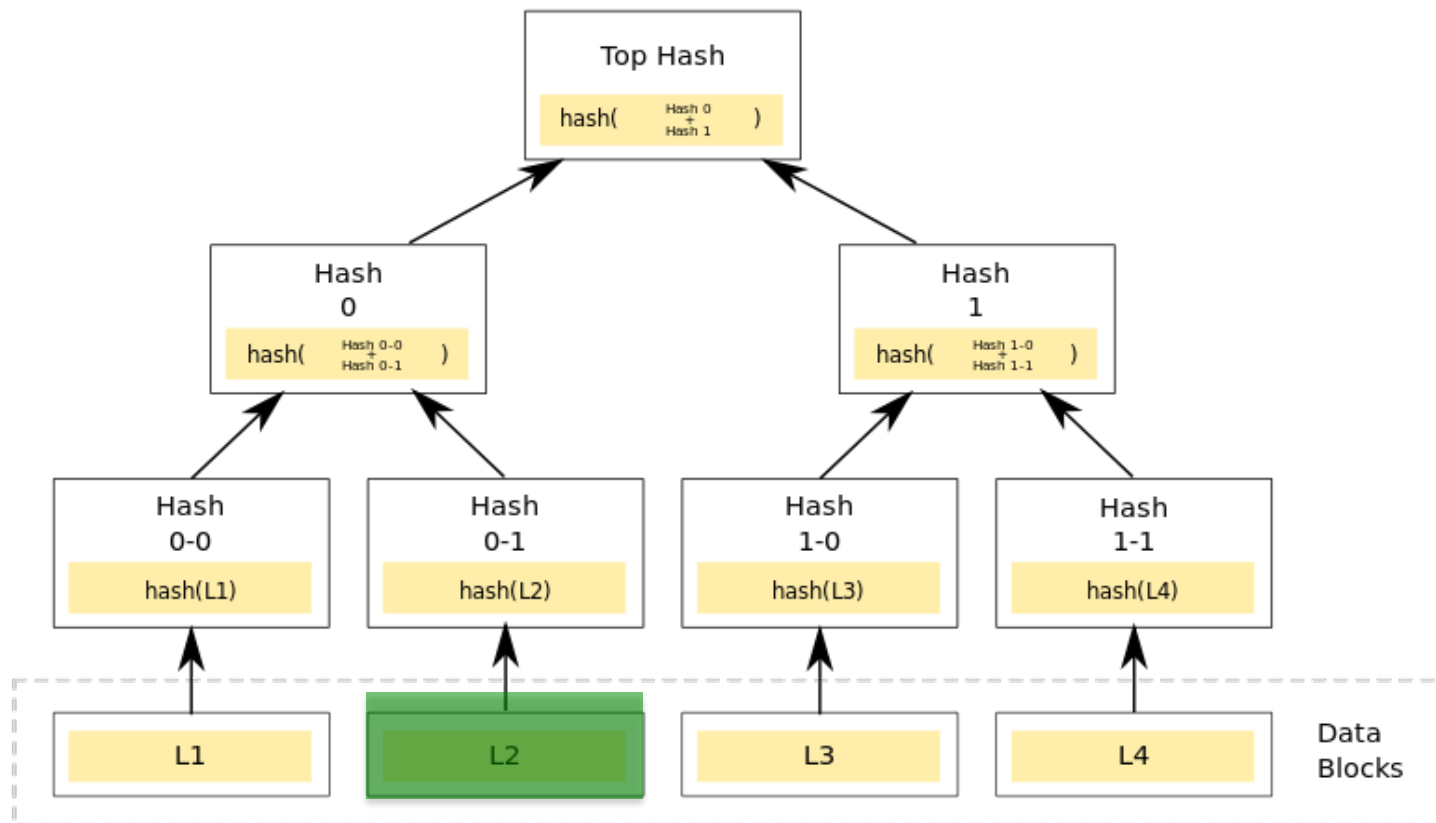


# Idea : Hash Tree ( Merkle Tree )

[https://en.wikipedia.org/wiki/Merkle\\_tree](https://en.wikipedia.org/wiki/Merkle_tree)

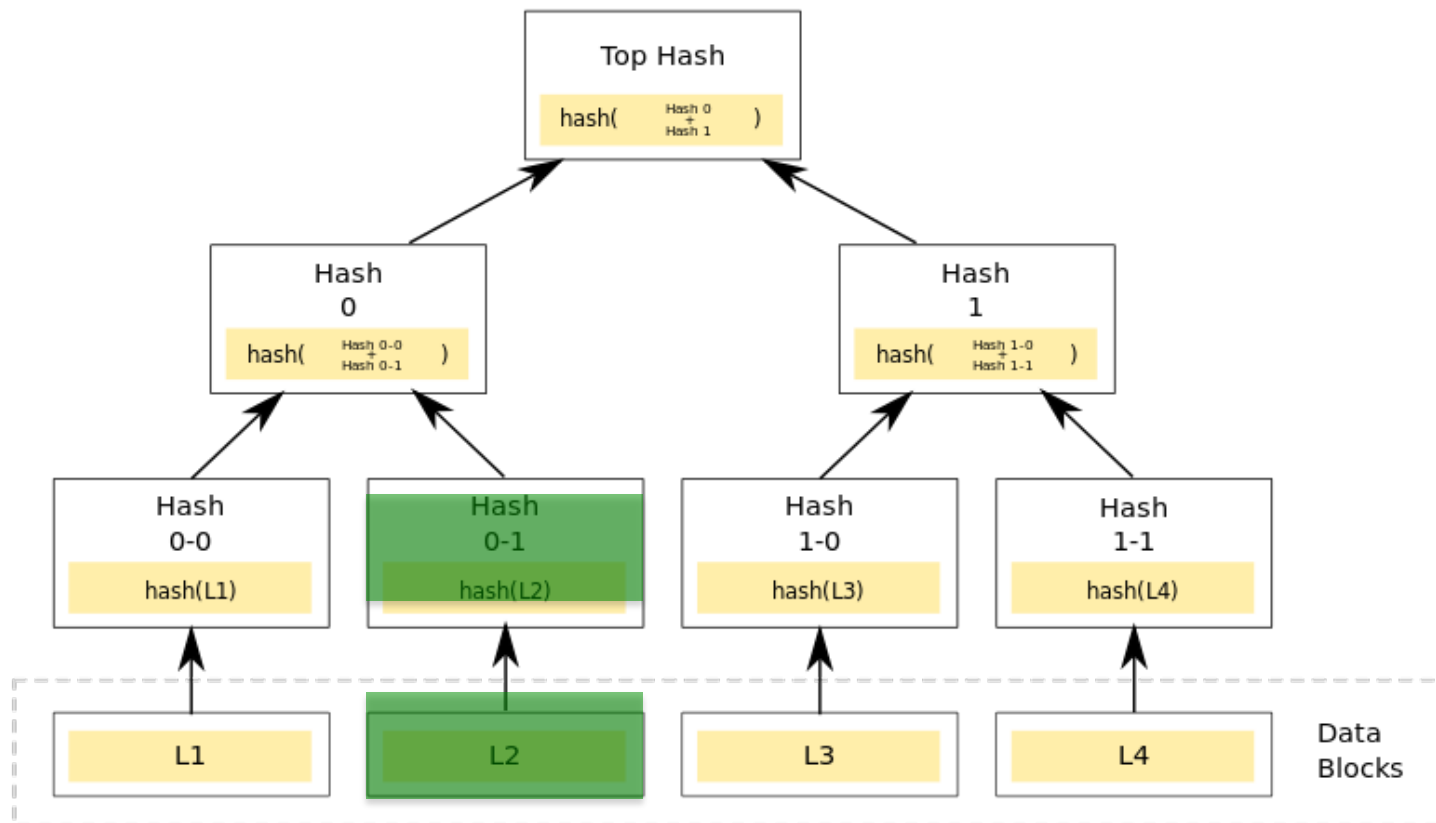


# Idea : Hash Tree ( Merkle Tree )

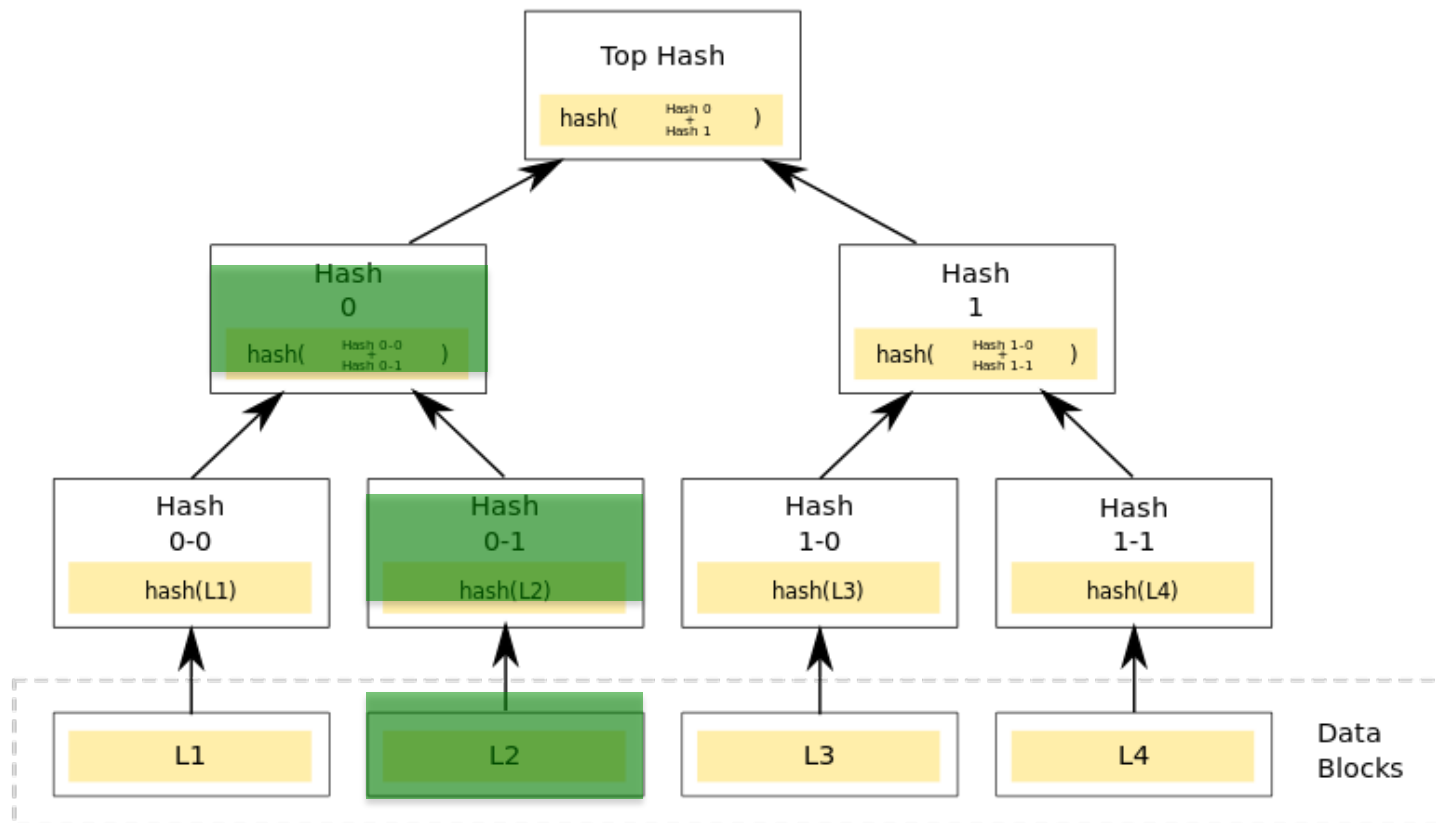




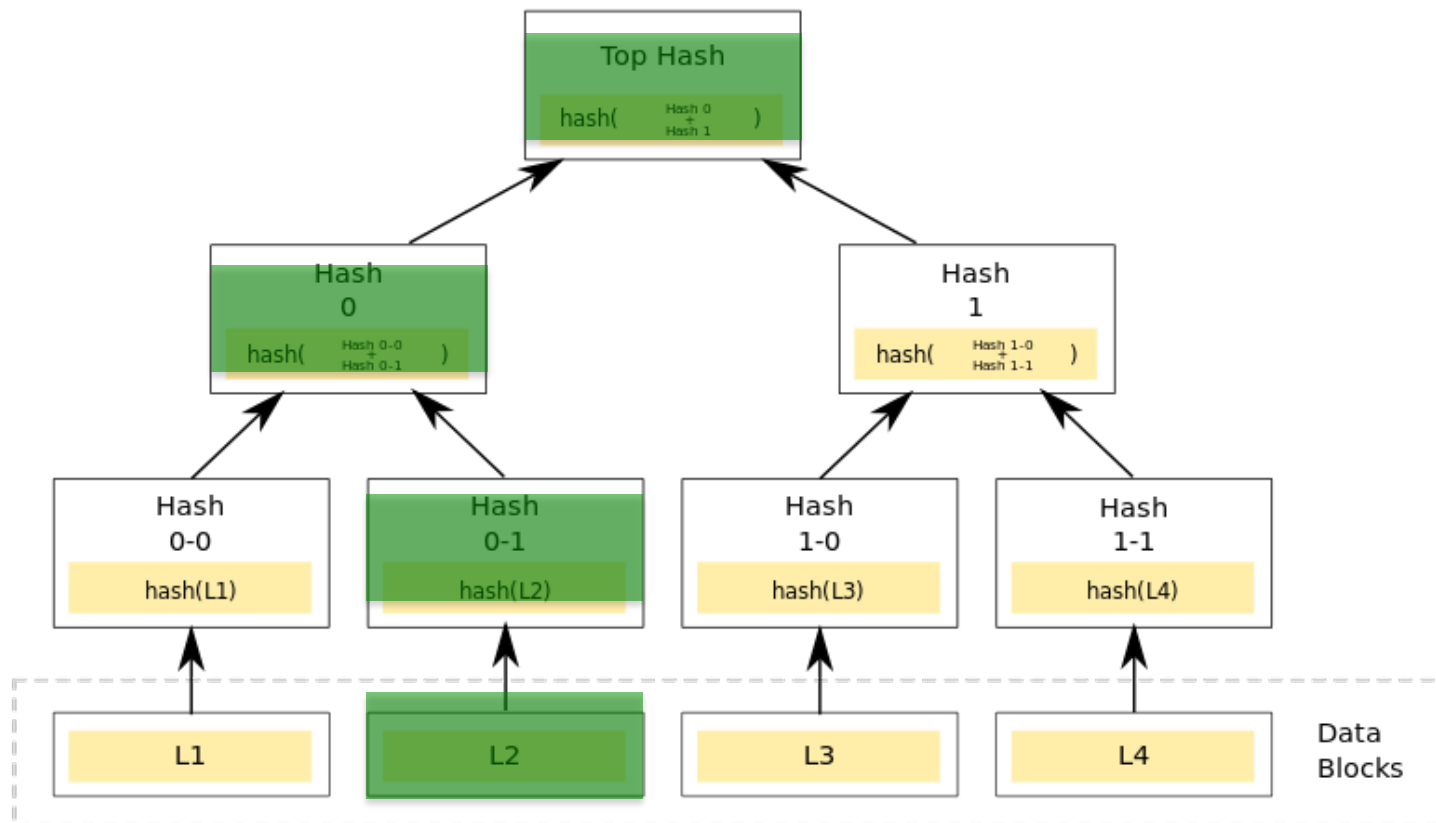
# Idea : Hash Tree ( Merkle Tree )



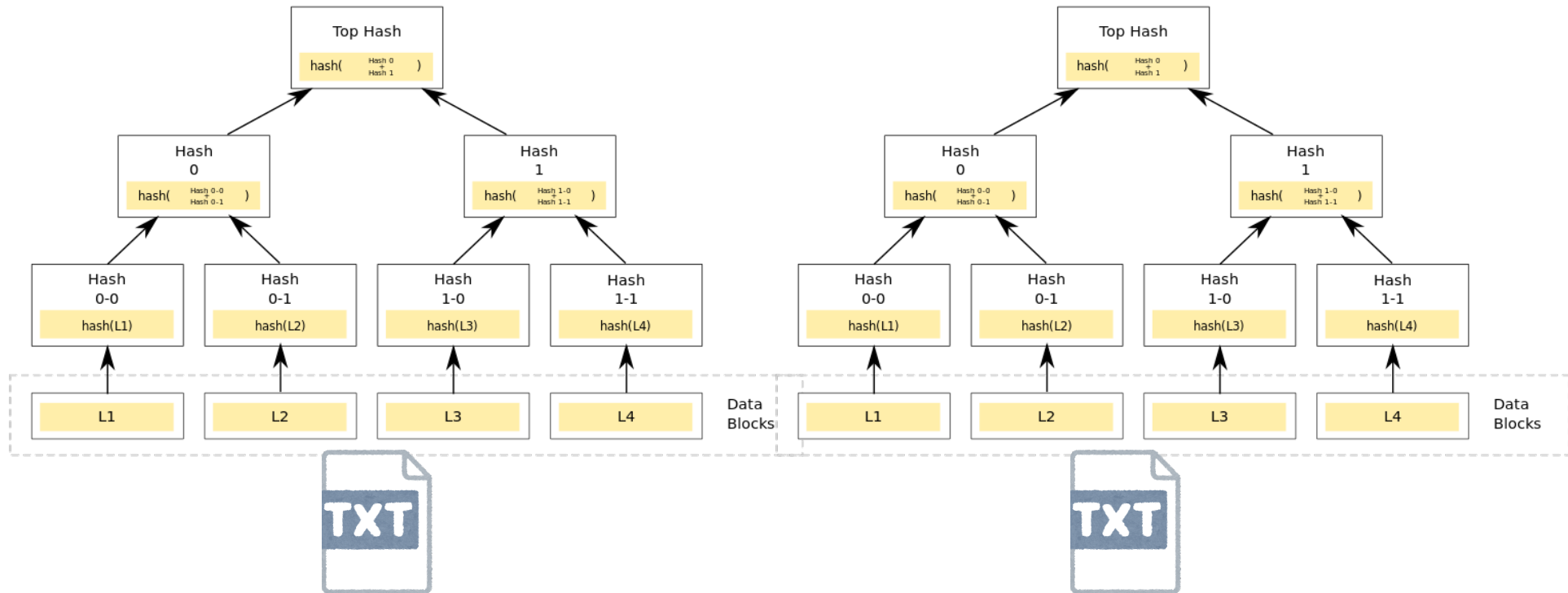
# Idea : Hash Tree ( Merkle Tree )



# Idea : Hash Tree ( Merkle Tree )

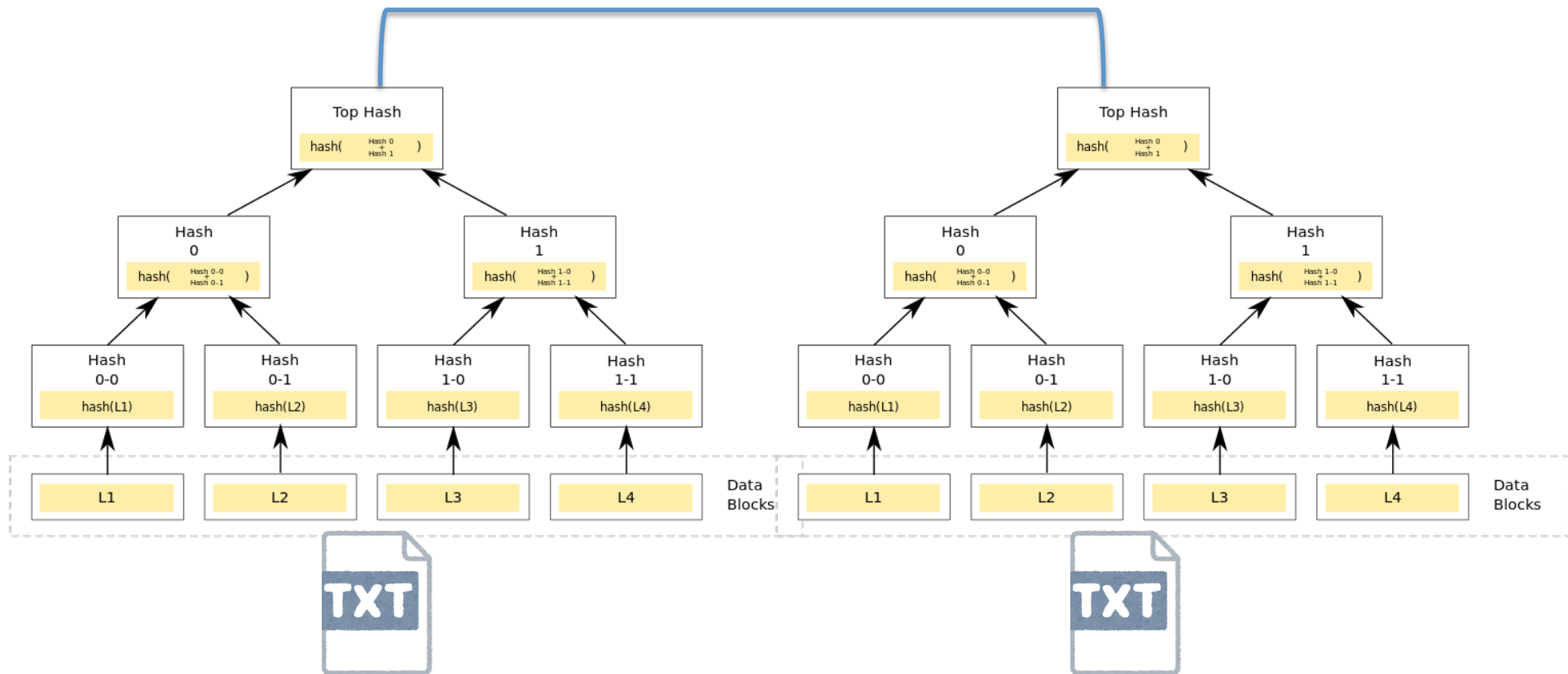


# Idea : Hash Tree ( Merkle Tree )

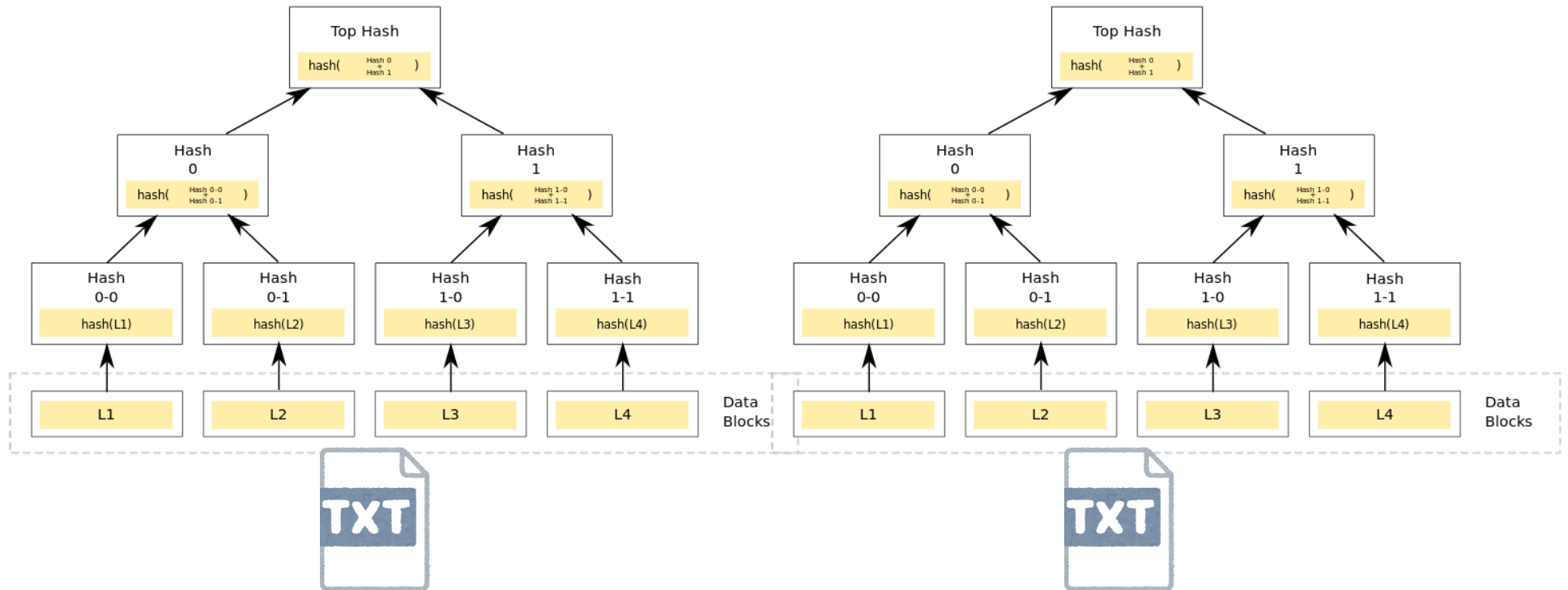


# Idea : Hash Tree ( Merkle Tree )

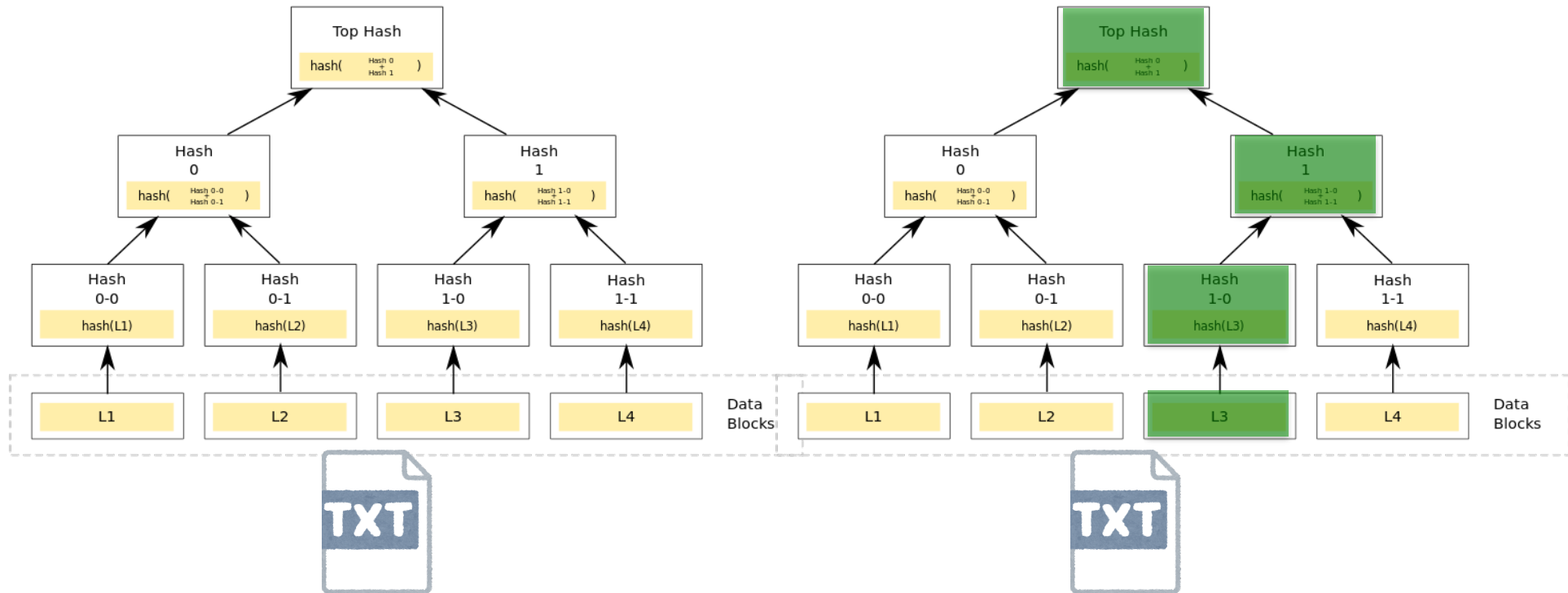
If the top hashes are same, 2 files' contents are same



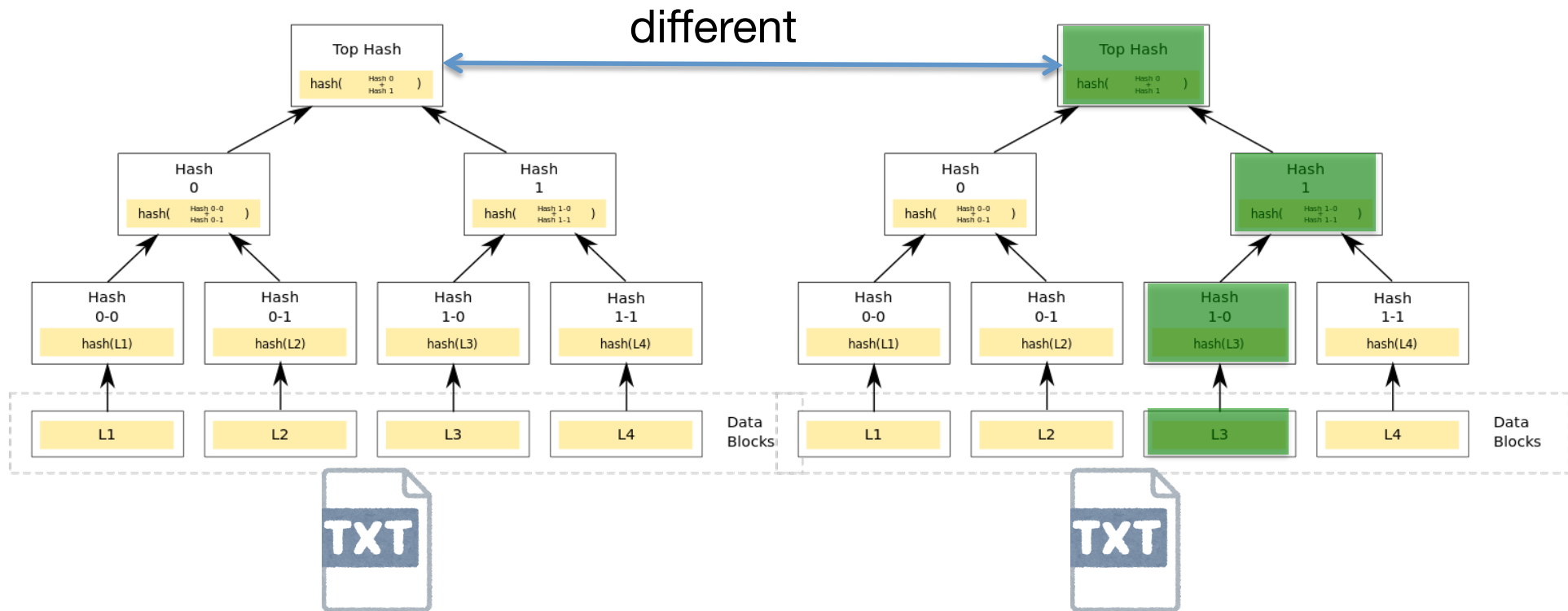
# Idea : Hash Tree ( Merkle Tree )



# Idea : Hash Tree ( Merkle Tree )

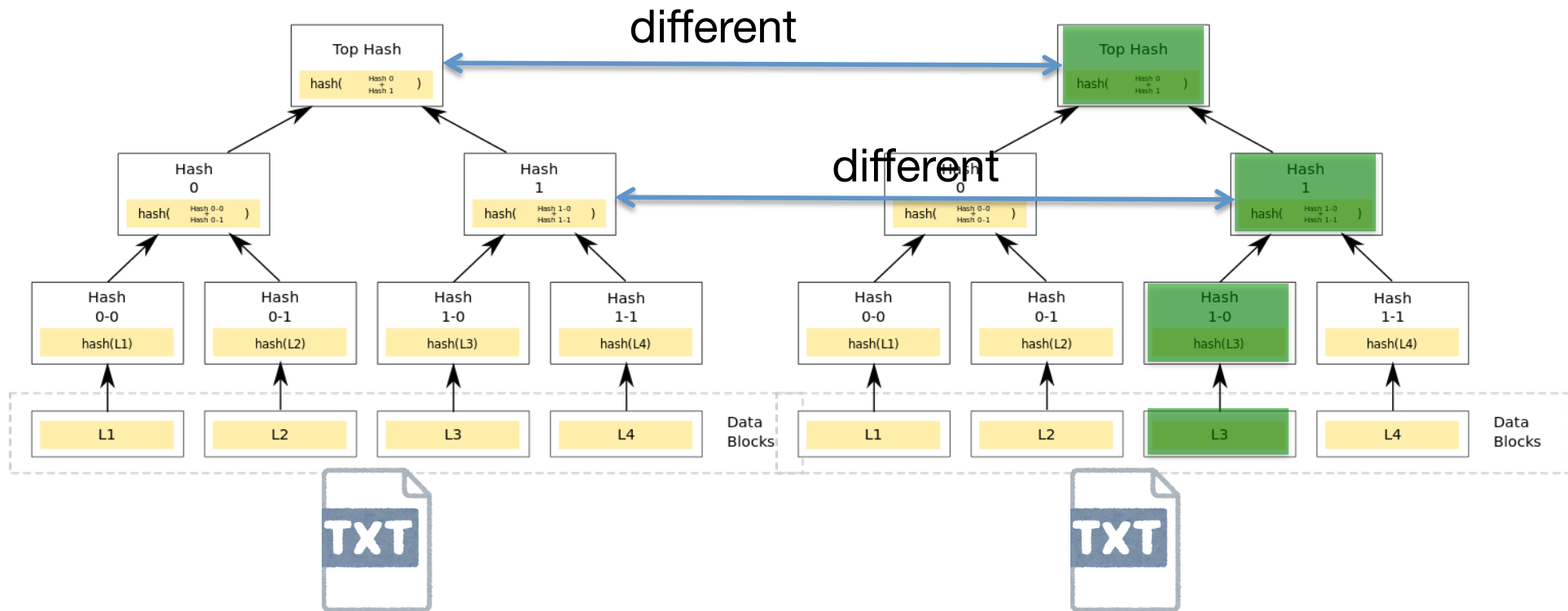


# Idea : Hash Tree ( Merkle Tree )

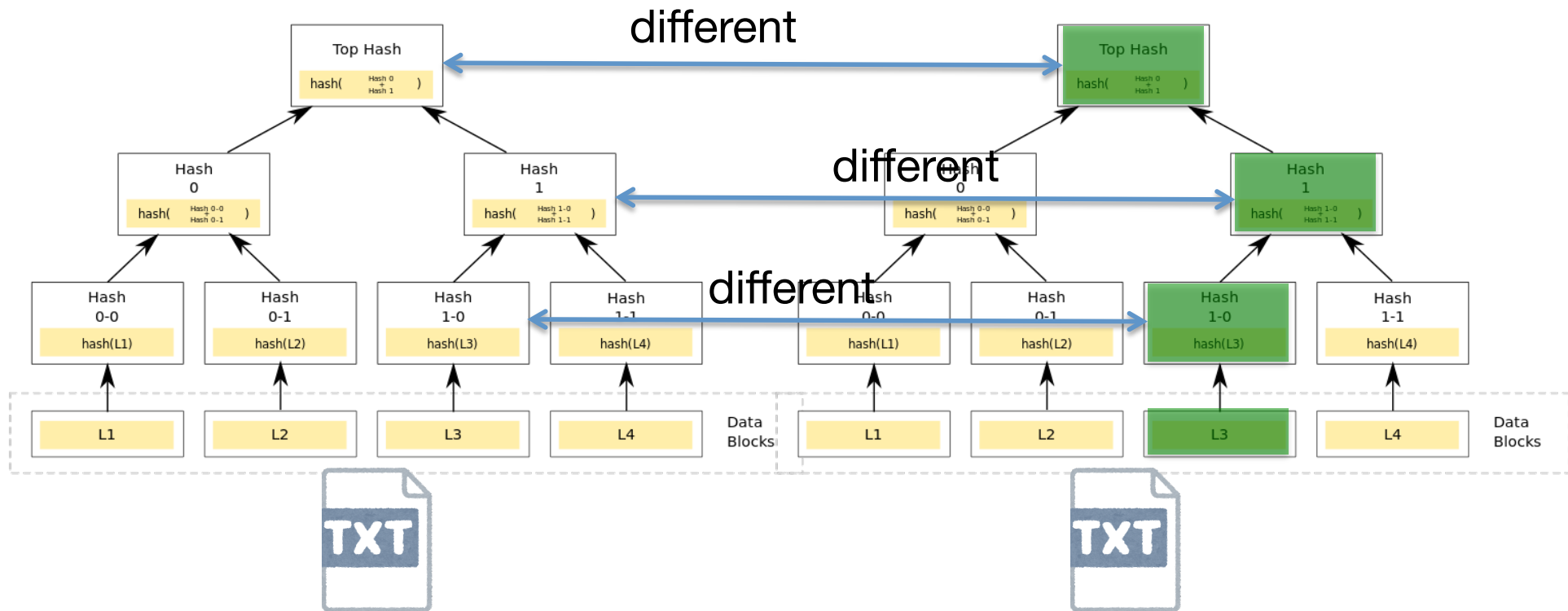




# Idea : Hash Tree ( Merkle Tree )

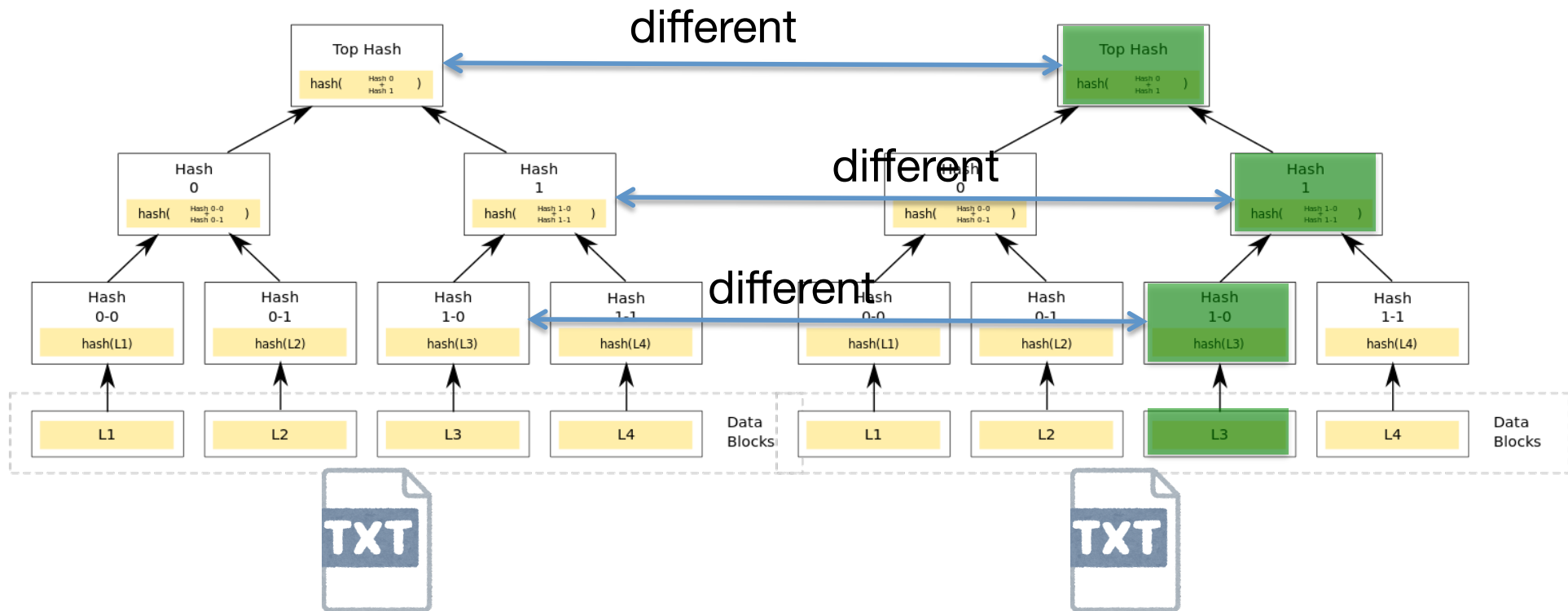


# Idea : Hash Tree ( Merkle Tree )



# Idea : Hash Tree ( Merkle Tree )

Block 3 is different!



## Idea : Hash Tree ( Merkle Tree )

- Instead of file contents, comparing hash trees to find out blocks which contain differences




- We do not need to read all file data from the disk



- Supposedly faster than default cmp

# Testing

- Generate a 100~ MB file named file1
- Run: `cp file1 file2`
- Modify 5~10 characters of file2



Please use  
the prepared tool

- Run: `sudo sh -c "echo 1 > /proc/sys/vm/drop_caches"`
- Run: `cmp -l file1 file2 > output_cmp.txt`
- Run: `YOUR_CMP -l file1 file2 > output_your_cmp.txt`
- Run: `diff output_cmp.txt output_your_cmp.txt`
  - If outputs of cmp and YOUR\_CMP are same, success

# Testing

- The ext4 file system
  - Only the name is modified so that we can avoid namespace conflict

<https://gitlab.montefiore.ulg.ac.be/yasukata/ext42>

- Test program

<https://gitlab.montefiore.ulg.ac.be/yasukata/diff-test>

## References ( maybe a lot more )

- Ext4 wiki
  - [https://ext4.wiki.kernel.org/index.php/Main\\_Page](https://ext4.wiki.kernel.org/index.php/Main_Page)
    - Howto : [https://ext4.wiki.kernel.org/index.php/Ext4\\_Howto](https://ext4.wiki.kernel.org/index.php/Ext4_Howto)
    - Disk Layout : [https://ext4.wiki.kernel.org/index.php/Ext4\\_Disk\\_Layout](https://ext4.wiki.kernel.org/index.php/Ext4_Disk_Layout)
- Book
  - Understanding the Linux Kernel 3<sup>rd</sup> Edition
    - by Marco Cesati, Daniel P. Bovet