43. Log-structured File Systems
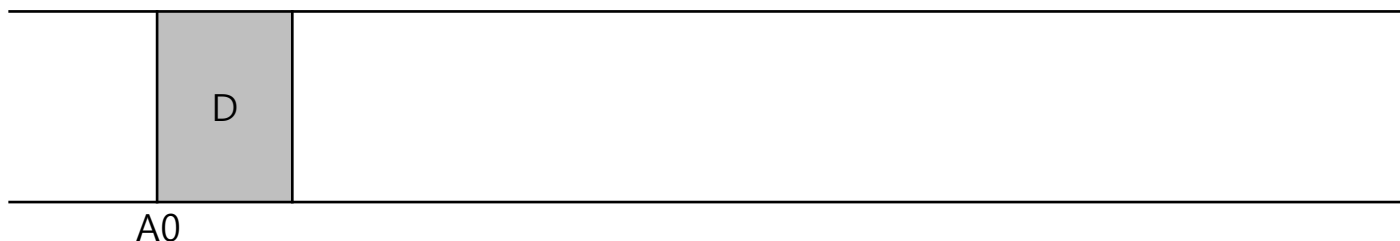
Operating System: Three Easy Pieces

LFS: Log-structured File System

- Proposed by Stanford back in 91

- Motivated by:

    - DRAM Memory sizes where growing.

    - Large gap between random IO and sequential IO performance (seek times)

    - Existing File System perform poorly on common workloads.

    - File System were not RAID-aware (small-write problem in RAID-4/5)

- Transform disk bandwidth into latency reduction!
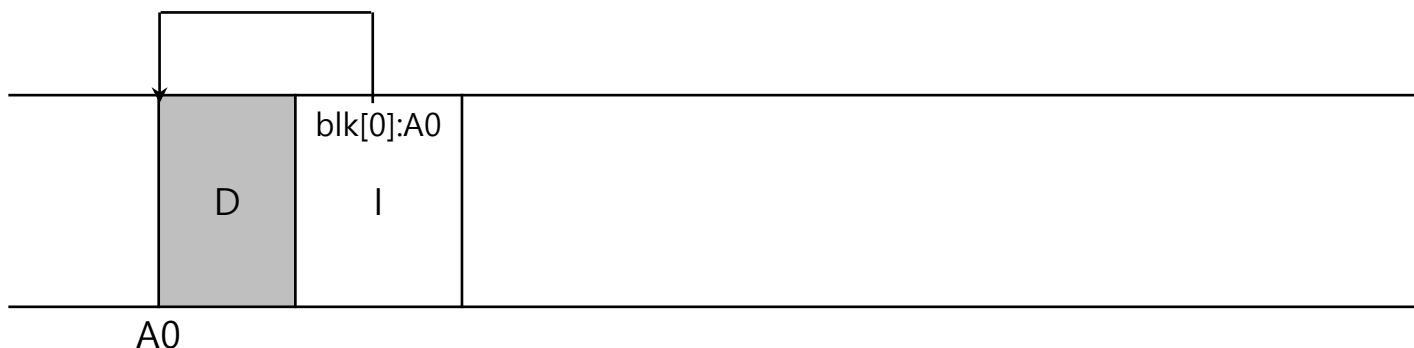
    - Transform random IO into sequential IO!

Writing to Disk Sequentially

▫ How do we transform all updates to file-system state into a series of sequntial writes to disk?

   ◆ data update

   ◆ metadata needs to be updated too. (Ex. inode)

- Writing single blocks sequentially does not guarantee efficient writes

  - After writing into A0, next write to A1 will be delayed by disk rotation

- Write buffering for effectiveness

  - Keeps track of updates in **memory buffer** (also called **segment**)

  - Writes them to disk all at once, when it has sufficient number of updates (or the user instruct so, i.e. call fsync)
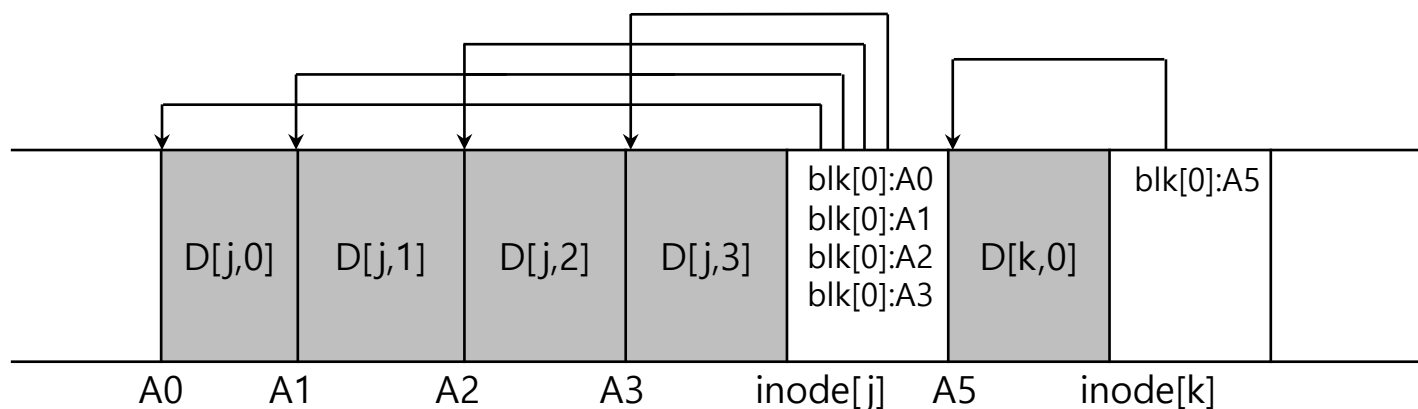
How Much to Buffer?

- ❑ Each write to disk has fixed overhead of positioning

  - ◆ Time to write out D MB

  $$T_{write} = T_{position} + \frac{D}{R_{peak}} \quad (43.1)$$

  ($T_{position}$: positioning time, R_{peak}: disk transfer rate in MB/s)

- ❑ To amortize the cost, how much should LFS buffer before writing?

  - ◆ Effective rate of writing can be denoted as follows

  $$R_{effecitve} = \frac{D}{T_{write}} = \frac{D}{T_{position+}\frac{D}{R_{peak}}} \ (43.2)$$

□ Assume that $R_{effecitve} = F{\times}R_{peak}$ (F: fraction of peak rate, 0 < F < 1), then

$$R_{effecitve} = \frac{D}{T_{position} + \frac{D}{R_{peak}}} = F{\times}R_{peak} \text{ (43.3)}$$

□ Solve for D

$$\text{D} = \frac{F}{1-F} \times R_{peak} \times T_{position} \text{ (43.6)}$$

□ If we want F to be 0.9 when $T_{position} = 10 msec$ and $R_{peak} = 100 MB/s,$ then $D = 9MB$ by the equation.

◆ Segment size should be 9MB at least.

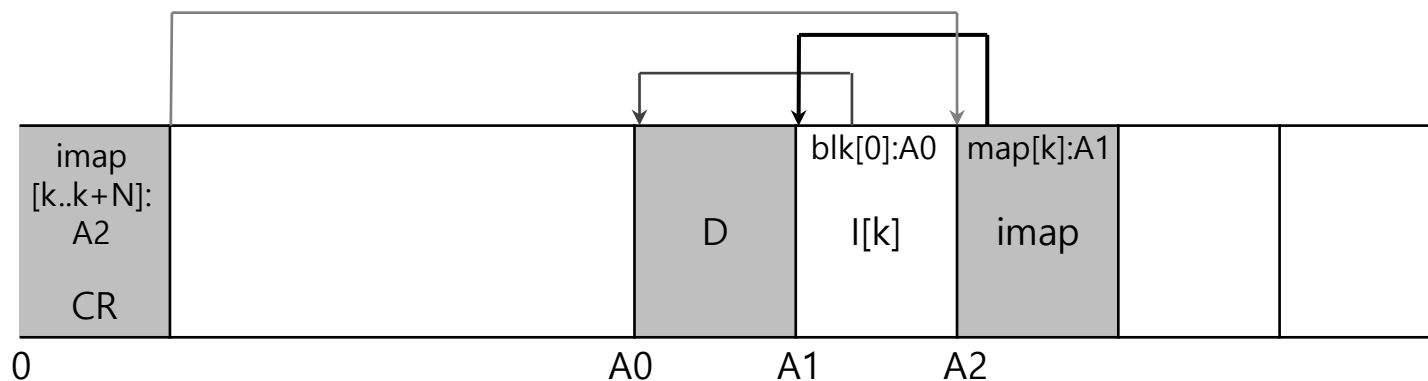- Inodes are scattered throughout the disk!

- Solution is through indirection "**Inode Map" (imap)**

- LFS place the chunks of the inode map right next to where it is writing all of the other new information (imap should be **persistent**)
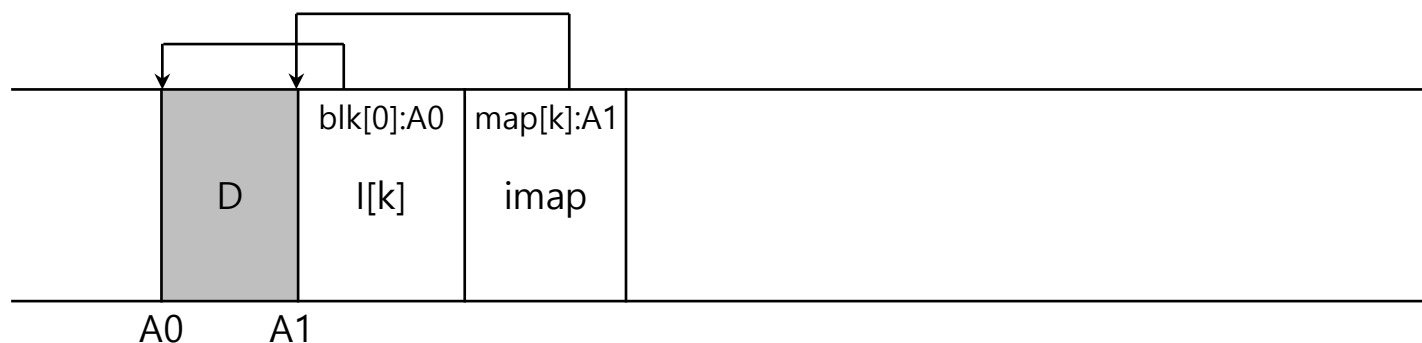
- But, how to **find** the imap?

The Checkpoint Region

- How to find the inode map, spread across the disk?

  - The LFS File system have fixed location on disk to begin a file lookup

- **Checkpoint Region** contains pointers to the latest of the inode map

  - Only updated periodically (ex. Every 30 seconds)
 - → performance is not ill-affected (like is with a "central" imap)

1. Read **checkpoint** region

2. Read entire **inode map** and cache it in memory

3. Read the most recent **inode**

4. Read a **block** from file by using direct or indirect or double-indirect pointers

- Directory structure of LFS is basically identical to classic UNIX file systems.

  - Directory is a file which data blocks consist of directory information

|	blk[0]:A0	(foo,k)	blk[0]:A2	map[k]:A1 map[dir]:A3			
D[k]	I[k]	D[dir]	I[dir]	imap			

A0   A1   A2   A3
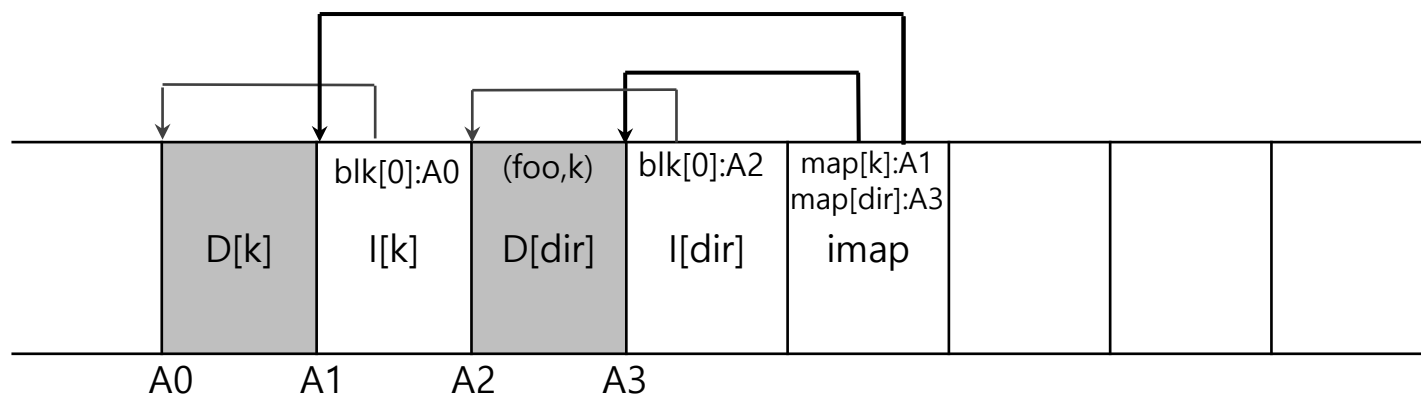
  - Directory avoids **recursive update problem** using imap (not inodes)
   - When a file change is location, directory won't be updated

Garbage Collection

- LFS keeps writing newer version of file to new locations.

- Thus, LFS leaves the older versions of file structures all over the disk, call as garbage.

- For a file with a singe data block (imap,dir, ... omitted by for simplicity)

  - Overwrite the data block: both old data block and inode become garbage

A0(both garbage)   A4
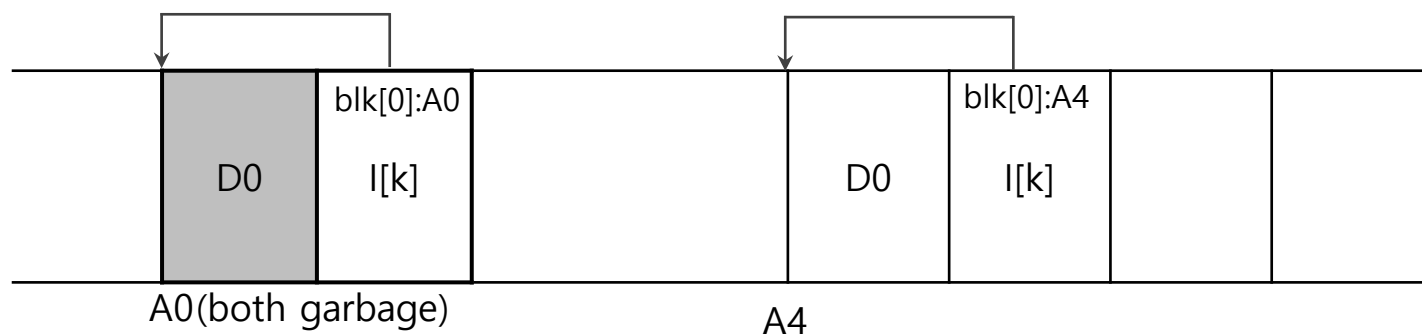
  - Append a block to that original file k: old inode becomes garbage

A0   (garbage)   A4
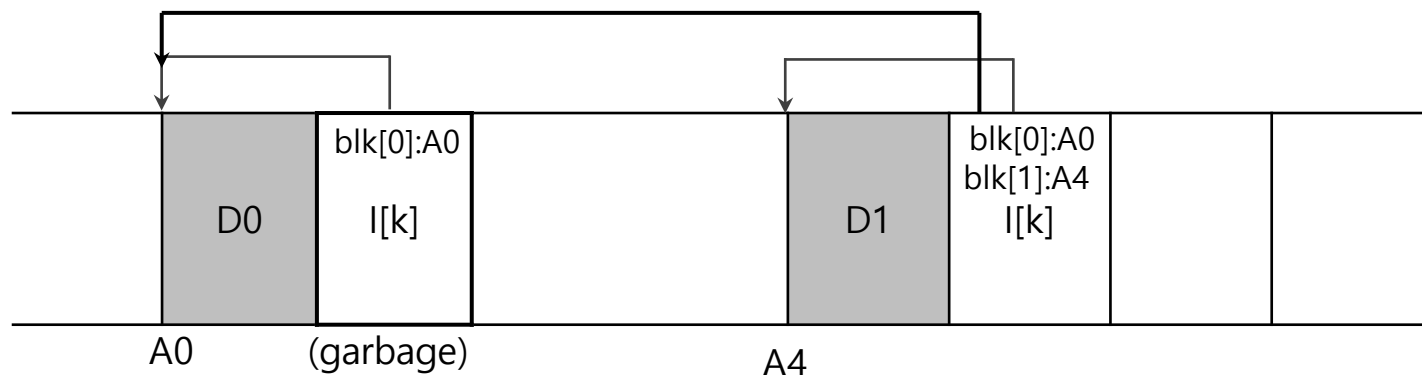
Handling older versions of inodes and data blocks

- One possibility: **Versioning file system**

  - keep the older versions around

  - Users can restore old file versions

- LFS approach: **Garbage Collection**

  - Keep only the latest live version and periodically clean old dead versions

  - **Segment-by-segment** basis

    - Block-by-block basis cleaner eventually make free holes in random location
      → Writes can not be sequential anymore

- Can be **performance critical**

  - In some benchmarks, performance can be terrible (v.gr if garbage collection interferes with the underlying workload)
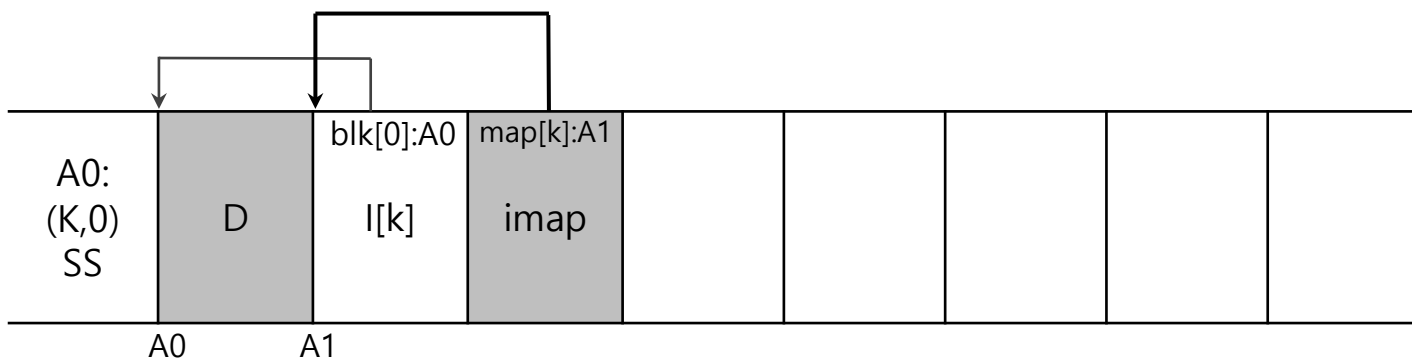
❑ **Segment summary block (SS)**

    ◆ Located in each segment

    ◆ Inode number and offset for each data block are recorded

❑ Determining Liveness

    ◆ The block is live if the latest inode indicates the block

A0: (K,0) SS	D	blk[0]:A0 I[k]	map[k]:A1 imap					

A0   A1

    ◆ **Version number** can be used for efficient liveness determining

```
//For a block D located in A with iNode N and offset T

(N,T) = SegmentSummary[A]

inode = Read(imap[N])

if ( inode[T] == A ){

    //Block D is alive }

else{

    //Block is dead (garbage)}
```

Which Blocks to Clean, and When?

❑ **When to clean**

  ◆ Periodically

  ◆ During idle time

  ◆ When the disk is full

❑ **Which blocks to clean**

  ◆ Segregate hot/cold segments

    ○ **Hot segment**: frequently over-written

      → more blocks are getting over-written if we wait a long time before cleaning

    ○ **Cold segment**: relatively stable

      → May have a few dead blocks, but the other blocks are stable

  ◆ Clean cold segment sooner and hot segment later

- ❑ Log organization in LFS

  - ◆ Checkpoint Region (CR) points to a **head** and **tail** segment

  - ◆ Each segment points to **next** segment

- ❑ LFS can easily recover by simply **reading latest valid CR**

  - ◆ The latest consistent snapshot may be quite old

- ❑ To ensuring atomicity of CR update

  - ◆ Keep **two CRs**, guarded by timestamps (TS) and use them alternatively

  - ◆ CR update protocol: TS➔ CR ➔ TS to **detect crashes** on CR update

- ❑ Roll forward (DB technique) to recover up to 30 secs

  - ◆ **Start from end of the log** (pointed by the latest CR with good TS)

  - ◆ **Read next segments** and adopt any valid updates to the file system

- Initially, garbage collection create some controversy and prevent the idea to success

- Copy-on-write is used by most state-of-the-art FS

  - ◆ BTRFS

  - ◆ ZFS

  - ◆ Some SSD FS

□ Disclaimer: This lecture slide set is used in AOS course at University of Cantabria by V.Puente. Was initially developed for Operating System course in Computer Science Dept. at Hanyang University. This lecture slide set is for OSTEP book written by Remzi and Andrea Arpaci-Dusseau (at University of Wisconsin)