CS 333 - LAB 7 - REPORT
140050037-140050048

1. The file system configurations that we are choosing here are ZFS with dedup on and off. We are trying to demonstrate how the total memory used varies when you have same data across different files with different configurations of file systems.

   IMPLEMENTATION details : The dedup feature is implemented as follows. It occurs when we try to write to disk. It implements dedup using checksums. When we are writing a block of data , and it has same checksum as a block already written, it is considered as a duplicate and just a pointer to that block is stored on disk. And the checksums are stored in a data structure called DDT (De-duplication table) for efficiency reasons.

2. **Workload**: I am creating a file and making a copy of it with a different file name.
   (i,e I am first creating a file with some random data with <filename1> as its name. Then I am making a copy of that file with the cp command and naming it <filename2> . So now I am having two different files with same data stored in them.
   I will do the above steps with deduplication on and off, and we will compare the results)

   Metrics: Overall disk usage

   Comparison of metrics:
   The original file <filename1> has size 100 MB
   With deduplication off:
   The overall disk usage for these two files is observed to be 201 MB , which is around twice the original file size, which is very much expected

   With deduplication on:
   The overall disk usage for these two files is observed to be 102 MB , which is approximately the same as the original file size, which is because both the files have same data, as dedup is turned on, it does not create a copy of the data blocks.

3. **DISADVANTAGE**: While the deduplication feature leads to lower disk usage, it may result in significantly higher CPU consumption. (This is because we have do all the checksum business, and compute the checksum for writing into DDT. This needs more CPU usage)
   With deduplication off:
   CPU usage is 3% approximately
   With deduplication on:
   CPU usage is 17% approximately
   (It shows up as a peak while you run the cp command in the CPU usage curve , which can be seen using system manager in ubuntu.)
   (I have attached the screenshots in the submission folder)

CS 333 - LAB 7 - REPORT

140050037-140050048

COMMANDS USED

To turn the dedup ON :

$ sudo zfs dedup=on datastore

(datastore is the name of the partition)

To turn the dedup OFF :

$ sudo zfs dedup=off datastore

To create a file with name <filename1>, 100 MB of random data :

$ sudo dd bs=1M count=100 if=/dev/urandom of=<filename1>

To see the statistics of disk usage and all:

$ sudo zpool list