abstract

We did a deduplication analysis on around 50TB compressed image dataset and found that we can save about half of the storage space by doing file-level deduplication after decompressing all the images. However, deduplication affacts Docker registrys performance especially for pulling performance because restoring layers require fetching files and compressing them into layers. To reduce deduplication restoring overhead(i.e., decompress/compression and additonal networking and I/O), we propose a novel architecture that intergrates compressed file-friendly deduplication and container user-friendly caching.