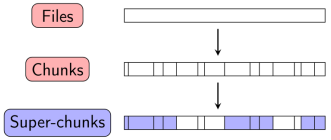
Super-chunking in Deduplication Storage Systems

Knowledge Lab

Birkbeck UNIVERSITY OF LONDON

Research Student
Jaybe Ammons

Supervisors
Trevor Fenner
David Weston



Birkbeck UNIVERSITY OF LONDON

Department of Computer Science and Information Systems

Research Aims

Our research is motivated by problems facing storage deduplication systems (DSS). We hope to devise mechanisms that reduce storage requirements. One technique we are investigating is a mechanism to **reduce metadata storage requirements**, especially for a dataset which is backed up periodically, **an evolving dataset.**

Research Methodology

A prototypical DSS, as described in [1], was implemented. Figure 1 shows the data flow in the system, followed by descriptions.

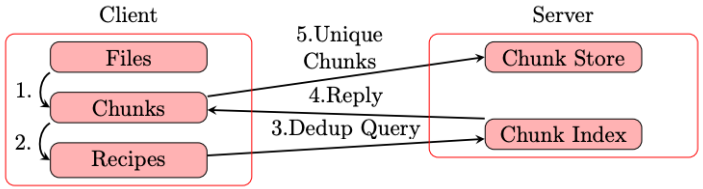


Figure 1. Prototypical DSS

1. Segment Files into Chunks, hash Chunk for Fingerprint
2. Build lists of Fingerprints making Recipes
3. Client sends Recipes to Server
4. Server returns list of Chunks not previously saved
5. Client packages up Chunks on list, sends to Server

Super-chunks are groups of consecutive chunks identified by a super-chunk fingerprint (hash of the chunk fingerprints). Figure 2 shows super-chunks being added to the prototypical system. We generated super-chunks from FSL Trace Data [2], and noted the differences in the volume of metadata.

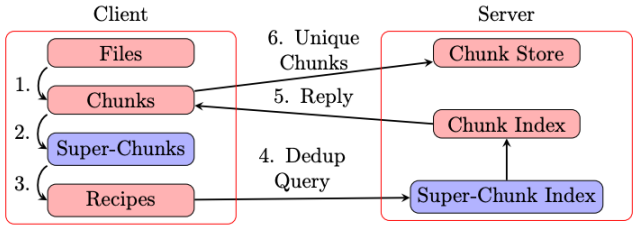


Figure 2. DSS with Super-chunking

Research Results

We found that recipe storage requirements without super-chunking were 91GB. Using super-chunking reduced the recipe size to 20GB, a reduction of 78%, as seen in Figure 3a.

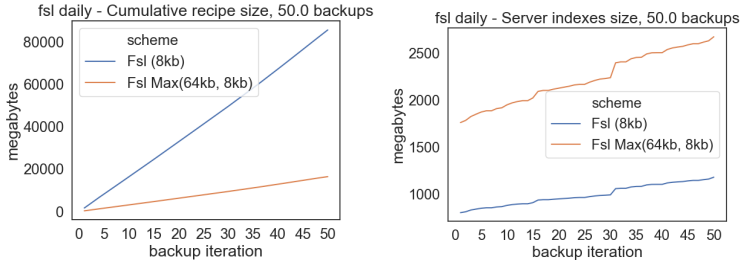


Figure 3. 50 FSL Trace backups: a) Cumulative Recipe Size, **b)** Server Index Size

This reduction in recipe size by 71GB does not come without cost. An additional super-chunk index is required on the server. This index required an additional 1.5GB of server space, as seen in Figure 3b.

Future Research

The super-chunk index size in Figure 3b (orange line) increases faster than the chunk index size (blue line). This may be caused by a lack of convergence of super-chunk boundaries, or inefficiencies in generating super-chunks with a mix of new and existing constituent chunks. This would be a good area for future research.

References

[1] Xia W, Jiang H, Feng D, Douglis F, Shilane P, Hua Y, et al. A Comprehensive Study of the Past, Present, and Future of Data Deduplication. Proceedings of the IEEE, vol. 104, no. 9, pp. 1681–1710, 2016.

[2] Sun Z, Kuenning G, Mandal S, Shilane P, Tarasov V, Xiao N, et al. A long-term user-centric analysis of deduplication patterns. 32nd Symposium on Mass Storage Systems and Technologies (MSST), pp. 1-7, 2016.