DADA: Duplication Aware Disk Array

Yiying Zhang* and Vijayan Prabhakaran†
University of Wisconsin-Madison and Microsoft Research†*

Primary storage systems contain only moderate degree of duplication and do not benefit greatly from deduplication. Instead of removing duplicates, we propose a duplication aware disk array (*DADA*), which keeps track of block duplication and uses the duplicate contents to improve the reliability and availability of disk arrays on primary storage systems.

Problem: Duplicate contents on disk have received a lot of attention in the form of deduplication storage systems [1–3], which are mainly used for archival and backup storage. In contrary, primary storage systems such as file servers and web servers have much less duplication and are accessed more randomly and more frequently. Duplication within primary storage systems thus has received fewer attention [4]. However, from our analysis of 5 different file server contents, we found moderate duplication (21-28%) and few content role changes (i.e. a block changing from or into being a duplicate).

Solution: Instead of removing duplicates like a deduplication storage system, we propose a duplication aware disk array (*DADA*), which recognizes and keeps duplicate contents on disks and make use of its knowledge of duplicates to improve the reliability and availability of disk arrays in the following three ways.

First, *DADA* can be used to reduce the time spent on disk scrubbing to detect latent sector errors (LSE). *DADA* scrubs only one representative block for a set of duplicates, which can be used to either reduce the amount of disk background activities or to permit more aggressive scrubbing and have a more reliable disk array.

Second, *DADA* reduces RAID reconstruction time during recovery. To reconstruct a disk, *DADA* first recovers only the unique block contents. It then starts user-level applications and in the same time recovers the rest of the disk. From application-level perception, recovery time is reduced and thus overall system availability is increased.

Finally, *DADA* provide better reliability than a default RAID array even though it does not scrub duplicated blocks. Using the knowledge of duplicated disk blocks, *DADA* can recover from certain latent sector error failures that are unrecoverable with default RAID.

There are two key challenges in designing *DADA*. First, duplicates may be scattered all over a disk causing disk seeks when *DADA* skips them during scrubbing or recovery. How we choose representative blocks for duplicates affects the amount of disk seeks. We propose several algorithms of choosing representative blocks so that duplicated blocks can be skipped with fewer disk seeks. These algorithms include simple ones such as randomly picking a representative block or picking one with the lowest offset, and more sophisticated ones which look at the regions duplicated blocks are in and choose respresentatives block in the region with most distinct contents. *DADA* may further remap duplicates, but only in regions with highest fragmented duplicates. Second, since duplicates are not scrubbed, they are subject to undetected LSEs. We analyzed all different failure senarios due to LSEs and demonstrate how *DADA* uses its knowledge of duplicates to recover from LSEs on unscrubbed blocks and how it can recover from more failures than a default RAID.

*Yiying Zhang is a student and will be presenting the poster.

References

[1] N. Jain, M. Dahlin, and R. Tewari, "TAPER: Tiered Approach for Eliminating Redundancy in Replica Synchronization," in *Proceedings of the File and Storage Technologies Conference*, (San Francisco, California), pp. 281–294, December 2005.

[2] S. Quinlan and S. Dorward, "Venti: A New Approach to Archival Storage," in *Proceedings of the File and Storage Technologies Conference*, (Monterey, California), pp. 89–101, January 2002.

[3] B. Zhu, K. Li, and R. H. Patterson, "Avoiding the Disk Bottleneck in the Data Domain Deduplication File System," in *Proceedings of the File and Storage Technologies Conference*, (San Jose, CA), pp. 269–282, Feb. 2008.

[4] E. Kruus, C. Ungureanu, and C. Dubnicki, "Bimodal Content Defined Chunking for Backup Streams," in *Proceedings of the File and Storage Technologies Conference*, (San Jose, CA), pp. 239–252, Feb. 2010.