Retrospective: Avoiding the Disk Bottleneck in the Data Domain Deduplication File System

Kai Li ✉ 🏠

Department of Computer Science, Princeton University, NJ, USA

Abstract

The paper titled "Avoiding the Disk Bottleneck in the Data Domain Deduplication File System" [3] describes several fundamental ideas behind the file system that drives Data Domain's deduplication storage products. Initially submitted to the 2007 ACM SIGOPS Symposium on Operating System Principles (SOSP), the paper was rejected by its program committee. It was subsequently submitted and accepted for publication at the USENIX Conference on File And Storage Technologies (FAST) in 2008. Twelve years later, it was honored with the USENIX Test-of-Time Award. This retrospective explores the paper's historical significance and impact, analyzes the reasons behind its initial rejection, and suggests methods to enhance the paper review process in the academic community.

2012 ACM Subject Classification Hardware → Communication hardware, interfaces and storage

Keywords and phrases Deduplication, file systems, compression

Digital Object Identifier 10.4230/LIPIcs.FUN.2024.33

Category Salon des Refusés

1 Innovation History

Founded in October 2001, Data Domain, Inc. aimed to replace traditional tape libraries in data centers with advanced disk-based storage products for backup and disaster recovery purposes. This vision was inspired by similar transitions happening at the time in the consumer electronics sector, where MP3 players (such as Apple's iPod) and Digital Video Recorders (DVRs) were supplanting music cassettes and VCR tapes, respectively. Both leveraged lossy compression methods (MP3 and MP4), achieving compression ratios of at least 10X, thus rendering disk-based devices economically competitive with their tape-based counterparts.

There were two significant challenges in developing disk-based storage systems to supplant tape libraries in data centers. The first challenge involved developing a compression method capable of achieving a 10X compression ratio to make disk-based systems economically economically competitive with similar capacity tape libraries. However, unlike the lossy compressions for audio and video data used in consumer products, the compression needed for this purpose had to be lossless. Traditional lossless compression methods, such as the Ziv-Lempel algorithm [4], typically only managed 2-3X compression, which varied according to the nature of the data.

The second challenge was to ensure high deduplication throughput while maintaining low costs. This was necessary to ensure that backups could be completed within a limited window of a few hours, thereby preventing any disruption to the normal daytime operations of the primary storage systems. Achieving this balance of high-throughput and cost-effectiveness was imperative. Although there were proposals about deduplication file systems [2, 1], none of them addressed this challenge.

In January 2002, we developed the key ideas for a high-throughput deduplication system designed to overcome both major challenges. We quickly implemented these strategies into a basic prototype and conducted tests using production backup data from three data centers.

© Kai Li;
licensed under Creative Commons License CC-BY 4.0

The encouraging results from these tests validated our ideas to create a high-throughput deduplication file system on cost-effective, standard hardware, achieving 20-30X lossless compression for backup data.

The technology validation convinced us to move forward with the design and development of a deduplication storage product. Additionally, we designed a data protection ecosystem that utilized these deduplication storage systems for storing local backups and efficiently transferred deduplicated backup data to the cloud for disaster recovery, eliminating the need for tape libraries in data centers and transporting tapes to remote sites. By 2004, we had successfully completed and released the products.

2 Paper Submission, Rejection and Publication

After five years of developing and successfully launching three generations of products in the market, we decided to document and share some of the key ideas behind the Data Domain file system. We submitted our paper to the ACM SIGOPS Symposium on Operating System Principles (SOSP) in 2007.

Regrettably, the SOSP program committee did not accept our submission, citing a lack of detailed evaluations as the primary reason. This decision came as a surprise to us, given that our deduplication file system was considered state-of-the-art in the storage industry, and our product line had generated over $100 million in revenue with over an 80% gross margin that year.

The reviewers did not recognize that their expected detailed comparative evaluations would require significant modifications to the file system, an unrealistic demand for a complex and sophisticated industry product. Therefore, after making minor edits, we resubmitted our paper to the USENIX Conference on File and Storage Technologies (FAST), where it was published in February 2008 [3].

3 Impacts

The Data Domain product line, powered by the deduplication file system described in our paper, dominated the backup storage market, capturing over 65% of the market share since its launch. Its revenue soared to $570 million in 2009 and surpassed $1 billion in 2010, supplanting the traditional tape libraries in data centers. The system's efficient high-compression ratio enabled the product line to sustain a gross margin exceeding 80%.

In the academic sphere, the paper has been extensively cited within the storage systems research community. In recognition of its lasting influence, the paper was awarded the USENIX Test-of-Time Award at the FAST Conference in 2020[1].

4 What's Not Included

The paper does not cover several key components of the deduplication file system:

- A concurrent GC (Garbage Collection) component to reclaim storage space of deleted data. The physical space of a data segment can be reclaimed only when it is not used by any file. The challenge is to accomplish this on-disk garbage collection using a small amount of memory and to keep up with the high deduplication throughput.

[1] https://www.usenix.org/conferences/test-of-tiemsme-awards

Concurrent Garbage Collection (GC): This component is essential for reclaiming storage space from deleted data. It only frees up the physical space of a data segment when it is no longer used by any file. On average, a data segment is shared by over 10 files in a deduplication file system. The challenge lies in performing this on-disk garbage collection efficiently using minimal memory while maintaining high deduplication throughput.

- Physical Data Replication: This component handles the replication of physical data containers over the Internet without the need to rebuild metadata structures remotely. It is designed for high-throughput, 1-to-1 data replications across high-speed network links.
- Logical Data Replication: Unlike physical data replication, this file-level protocol transfers individual physical segments. It is particularly useful for many-to-1 data replications to prevent the transfer of duplicated data segments at the destination, significantly reducing network bandwidth requirements.
- Software RAID: This component implements an abstraction layer for block storage, ensuring reliability in the event of disk failures, power outages, OS software malfunctions, and during the replacement of failed disks.
- Error Detection and Correction: This component regularly scans the storage space to identify and correct data corruptions using stored error codes. Because corruption in a single data segment could compromise many files, its role is essential for maintaining data integrity.

To fully comprehend the Data Domain deduplication file system, one must be familiar with these components. It would have been advantageous for the community to read papers about them.

5 Reflections

Why was a high-impact paper rejected by a reputable conference? Several hypotheses can be considered:

- Expertise of Reviewers: The reviewers may not have been experts in storage systems, potentially lacking an understanding of the significance of advancements in the field.
- Expectations for Evaluations: Academic reviewers are trained to expect detailed comparative evaluations, feasible with a simulator but unrealistic with a complex industry product.
- Credibility of Claims: Reviewers could have found it difficult to believe that Data Domain's deduplication file system could achieve a 10X higher compression ratio and operate 10X faster than optimized, known compression tools.

Every program committee strives to select the best papers for its conference and aims to avoid overlooking high-impact submissions. To mitigate these issues, program committees could ensure that reviewers are chosen for their expertise in the relevant subject matter and that they have realistic expectations for evaluations of industry products. Additionally, a blend of anonymous and open review processes might help the committee better understand the credibility of the systems discussed in the papers during their final deliberations.

How can we determine if a systems paper has made a significant impact? Alan Perlis once remarked, "The proof of a system's value is its existence."[2] This insightful quote underscores that the real-world application and longevity of a system attest to its value and impact.

[2] http://www.cs.yale.edu/homes/perlis-alan/quotes.html

References

1 Athicha Muthitacharoen, Benjie Chen, and David Mazieres. A low-bandwidth network file system. In *Proceedings of the eighteenth ACM symposium on Operating systems principles*, pages 174–187, 2001.

2 Sean Quinlan and Sean Dorward. Venti: A new approach to archival data storage. In *Conference on File and Storage Technologies (FAST 02)*, Monterey, CA, January 2002. USENIX Association. URL: `https://www.usenix.org/conference/fast-02/venti-new-approach-archival-data-storage`.

3 Benjamin Zhu, Kai Li, and R Hugo Patterson. Avoiding the disk bottleneck in the data domain deduplication file system. In *Fast*, volume 8, pages 1–14, 2008.

4 Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, 24(5):530–536, 1978.