DEDIS: Distributed Exact Deduplication for Primary Storage Infrastructures

J. Paulo and J. Pereira
High-Assurance Software Laboratory, INESC TEC & Universidade do Minho
{jtpaulo,jop}@di.uminho.pt

1 Motivation and Challenges

Deduplication is now widely accepted as an efficient technique for reducing storage costs at the expense of some processing overhead, being increasingly sought in primary storage systems [7, 8] and cloud computing infrastructures holding Virtual Machine (VM) volumes [2, 1, 5]. Besides a large number of duplicates that can be found across static VM images [3], dynamic general purpose data from VM volumes allows space savings from 58% up to 80% if deduplicated in a cluster-wide fashion [1, 4]. However, some of these volumes persist latency sensitive data which limits the overhead that can be incurred in I/O operations. Therefore, this problem must be addressed by a cluster-wide distributed deduplication system for such primary storage volumes.

Although considerable space savings are obtainable, storage latency is critical in primary workloads so, deduplication must introduce negligible overhead to be viable. Traditional *in-line* deduplication includes computation inside the storage write path, adding unacceptable overhead in the latency of primary storage writes [6]. This penalty can be reduced by exploring data locality, however, it is only viable for specific storage workloads [5, 7]. On the other hand, *off-line* deduplication decouples aliasing from storage requests, reducing the latency penalty, but requiring additional temporary storage space and increasing the concurrency in storage accesses. In fact, given the overhead of copy-on-write mechanisms needed to avoid corrupting aliased data [1], even off-line deduplication must be confined to off-peak periods in order not to degrade latency. Off-peak periods in cloud infrastructures may be scarce, therefore deduplication should run continuously to detect duplicates promptly and maintain a small storage backlog. Finally, exact global deduplication across a cluster of storage servers is also challenging as it requires remotely accessing a global index, adding to the latency of both I/O and deduplication operations [2, 1].

SoCC'13, 1–3 Oct. 2013, Santa Clara, California, USA.
ACM 978-1-4503-2428-1. http://dx.doi.org/**10.114525236162528936**

2 DEDIS

We address the combined challenges of primary storage and global deduplication with DEDIS, a dependable and fully decentralized system that performs exact and cluster-wide off-line deduplication of primary volumes. Unlike previous systems, it works on top of any storage that exports an unsophisticated shared block device interface, and does not depend on data locality assumptions in the workload. Optimistic off-line deduplication is performed outside storage I/O requests that are intercepted and redirected to the correct storage address, at the fixed-size block granularity, by a layer that enables block aliasing. Deduplication is performed globally and exactly across the whole cluster by using a sharded and replicated fault tolerant distributed service that maintains both the index of unique block signatures and the metadata necessary for reference management and garbage collection. As other contributions, we leverage off-line deduplication to detect and avoid I/O hot spots, significantly reducing the number of costly copy-on-write operations. Also, batch processing and caching optimizations reduce the overhead introduced by DEDIS on I/O operations while increasing deduplication throughput. Finally, DEDIS tolerates hash collisions by performing byte comparison of blocks before aliasing them while imposing a minimal penalty in storage I/O latency. The resulting system can perform deduplication continuously in background without introducing significant overhead in foreground I/O applications.

Our experimental evaluation with an open-source prototype of DEDIS[1] in up to 20 nodes shows that deduplication scales and introduced less than 10% of overhead in primary storage latency.

[1] Open source available at: http://www.holeycow.org

References

[1] A. T. Clements, I. Ahmad, M. Vilayannur, and J. Li. Decentralized Deduplication in SAN Cluster File Systems. In *Proceedings of USENIX Annual Technical Conference (ATC)*, 2009.

[2] B. Hong and D. D. E. Long. Duplicate Data Elimination in a San File System. In *Proceedings of Conference on Mass Storage Systems (MSST)*, 2004.

[3] D. T. Meyer, G. Aggarwal, B. Cully, G. Lefebvre, M. J. Feeley, N. C. Hutchinson, and A. Warfield. Parallax: Virtual Disks for Virtual Machines. In *Proceedings of European Conference on Computer Systems (EuroSys)*, 2008.

[4] D. T. Meyer and W. J. Bolosky. A Study of Practical Deduplication. In *Proceedings of USENIX Conference on File and Storage Technologies (FAST)*, 2011.

[5] C.-H. Ng, M. Ma, T.-Y. Wong, P. P. C. Lee, and J. C. S. Lui. Live Deduplication Storage of Virtual Machine Images in an Open-Source Cloud. In *Proceedings of ACM/IFIP/USENIX International Middleware Conference*, 2011.

[6] S. Quinlan and S. Dorward. Venti: A New Approach to Archival Storage. In *Proceedings of USENIX Conference on File and Storage Technologies (FAST)*, 2002.

[7] K. Srinivasan, T. Bisson, G. Goodson, and K. Voruganti. iDedup: Latency-aware, Inline Data Deduplication for Primary Storage. In *Proceedings of USENIX Conference on File and Storage Technologies (FAST)*, 2012.

[8] J. Wright. Sun ZFS Storage Appliance Deduplication Design and Implementation Guidelines. http://www.oracle.com/technetwork/articles/servers-storage-admin/zfs-storage-deduplication-335298.html, 2011.