DELL Technologies

Reference: Review of Mr. Andrzej Jackowski's Thesis "Adapting Distributed Storage with Deduplication to Cloud Use Cases"

January 29, 2024

Dear Thesis Review Committee,

I was appointed to review Mr. Andrez Jackowski's doctoral thesis "Adapting Distributed Storage with Deduplication to Cloud Use Cases," and I have completed a thorough review. His thesis addresses three key topic areas. First, it presents a new object storage implementation with block level deduplication called ObjDedup. Second, it implements a novel technique, InftyDedup, to tier data to the cloud where deduplication computation takes place as well as efficient storage. Third, it addresses the problem of achieving high space utilization and failure resilience in a distributed storage system with heterogeneous storage with a technique called Derrick. All of these research topics were investigated with the highest level of thoroughness, and the implementation within HYDRAstor will be directly valuable to numerous storage customers. **I deem the thesis as sufficient to grant a Ph.D., and I further recommend it be awarded honorary distinction.**

In case you are not familiar with my background and ability to evaluate this thesis, I am a Distinguished Engineer at Dell Technologies, where I have worked since 2007 on deduplicated storage systems, compression, data characterization, and flash caching. I am currently architecting the next-generation data protection product to scale for increasing customer demands using a microservices architecture. I have invented over 160 US patents and have more than 40 publications in journals and conferences including Proceedings of the IEEE, ACM Transactions on Storage, USENIX Annual Technical Conference, and USENIX Conference on File and Storage Technologies. My publications and patents have been cited over 8,900 times. I co-chaired the 2021 ACM Workshop on Hot Topics in Storage and File Systems, regularly serve as a program committee member for conferences, and am currently an assistant editor for ACM Transactions on Storage. I have also shared datasets and tools with the larger research community, including storage traces used in Mr. Jackowski's thesis for deduplication studies. I received B.S. and M.S. degrees in computer science from Stanford University and M.A. and Ph.D. degrees in computer science from Princeton University. I am a Senior Member of the IEEE and was selected as the 2023 IEEE Philadelphia Section Engineer of the Year.

The early chapters of the thesis provide thorough background material about deduplicated storage, cloud storage, and distributed storage problems. The research overview in chapter 2 is itself a contribution to the research community since it provides an up-to-date, structured view of the history of deduplication research and succinct summary of hundreds of publications. Chapter 3 continues with the history of distributed storage, modern storage devices, cloud storage, and object storage interfaces. With this introduction to the field and problem areas, readers can place the following thesis research within the larger research context and appreciate its contributions.

Chapter 4 presents ObjDedup, which provides a new object storage interface to deduplicated storage, which was originally published in in IEEE Transactions on Parallel and Distributed Systems. The research begins with a study of 686 real-world deployments to understand the design requirements. The study showed that backup systems have more writes than deletes and 16% of cases showed capacity utilization above 80%, so efficient garbage collection is necessary. Metadata operations were identified as a new problem to solve since object storage tends to write objects that are orders of magnitude smaller than typical backup storage, which increases the metadata and operations on the metadata proportionally. The thesis presents new data structures and related algorithms called ObjectMetadataLog (OML) and ObjectMetadataTree (OMT). Updates are written to OML in memory to reduce latency and then applied to the OMT asynchronously. The OMT is a B+ tree-like structure handling most metadata. The OMT is itself distributed as SubOMTs based on hashes of the keys to workers. The distribution can be dynamically adjusted to avoid hot spots. Analysis of the algorithms proved that updates were related to the height of the OMT, which is at most 9 levels for very large clusters. The research was implemented within HYDRAstor, a commercially successful backup storage appliance with a distributed architecture. Experiments demonstrate that adding ObjDedup to HYDRAstor does not affect baseline write or read performance as the cluster grows, and resource overheads are reasonable. This results in a new protocol for deduplicated backup storage using object storage APIs.

Chapter 5 covers InftyDedup, which is an architecture for tiering from an on-premises deduplicated storage system to the cloud and leverages low-cost cloud services for both computation and storage. The work was originally published at the USENIX Conference on File and Storage Technologies. The design focuses on low-cost batch processing using spot instances and supports cross-system deduplication. Unlike previous work that relied on the on-premises appliance (i.e. local tier) for computation and therefore supported limited capacity in the cloud, InftyDedup has unlimited scalability because it runs independently in the cloud. The local-tier appliance uploads unprocessed file recipes (UFRs) consisting of a sequence of fingerprints representing a file, and the UFRs are batch processed periodically. Fingerprints are checked against a fingerprint index, and processed file recipes (PFRs) are generated. Any missing blocks are requested from the local tier and uploaded to the cloud. Garbage collection of deleted blocks was extended to migrate live blocks between hot and cold tiers based on a cost formula that incorporated file expiration dates and reference counts. Extensive experiments with the implementation in HYDRAstor and simulator showed the scalability of InftyDedup, its low-cost calculation design, and that cost-based assignment of blocks outperformed alternative designs. Even with highly varying read frequency, InftyDedup's design achieved excellent cost savings. Experiments with the real-world FSL traces were consistent with experiments with synthetic traces.

Derrick is presented in chapter 6 to address the problem of placing data blocks in a scale-out cluster. It was originally published in ACM Transactions on Storage. In large storage clusters, data is replicated (or encoded involving replicas) to protect against drive or full server failures. A group of data may have three replicas for example called components, and Derrick addresses the problem of where to place the components within the cluster to balance capacity usage and resilience requirements as well as approximately 20 other metrics. Derrick is subdivided into three algorithms. Central Balancing focuses on creating a data arrangement for a stable system that attempts to optimize the metrics. Transition Guide generates a transition plan to migrate data to the new arrangement while maintaining data availability and avoiding temporarily overloading the capacity of devices. Distributed Balancing is similar to Central Balancing but needs to respond quickly to device failures. All three algorithms use a variant of hill-climbing to incrementally improve a plan with the guidance of multiple heuristics. In experiments comparing to Swift and Ceph's CRUSH, Derrick more quickly generates better arrangements with the fewest data transfers. Derrick covers a complex systems optimization topic, and in places the clarity could be improved with a more structured presentation of how heuristics are implemented in the overall algorithms.

In conclusion, Mr. Jackowski's doctoral thesis is a very strong research effort that provides novel and useful solutions to multiple important problems in deduplicated and distributed storage. Implementing his research within HYDRAstor adds an additional level of complexity that most graduate students do not undertake and provides evidence that his research is of practical value for commercial uses beyond its theoretical contributions. His work is likely to influence future research and industry products. With the combination of research and product impact, I recommend it be awarded honorary distinction.

Regards,

Philip Shilane

Philip Shilane, Ph.D.
Distinguished Engineer at Dell Technologies