QoS by Reduction Ratio

Ken Kligmann

Pure Storage, Inc.

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Kligmann, Ken, "QoS by Reduction Ratio", Technical Disclosure Commons, (September 17, 2020)
https://www.tdcommons.org/dpubs_series/3614

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
Motivation

Users today may give special consideration when migrating highly dedupable workloads onto a storage array to avoid running out of space on the array. To address this, a user could throttle the write rate at the host level in order to help preserve inline deduplication as much as possible. This would specifically be done so that space on the storage array does not balloon during migration of heavily dedupable workloads.

For example, if a user has an array with 100TB usable capacity and 50TB of that space is already being used, there would be 50TB of usable capacity remaining. If the user has a workload where they expect to see a 5:1 data reduction ratio and that workload is 250TB in size, the workload should fit on the array consuming the remaining 50TB. If the user migrates the workload’s data to the array at a write rate above the array’s ability to preserve inline deduplication, the array could run out of space before the migration is complete, potentially impacting existing workloads.

When migrating large, highly dedupable workloads, one way to help ensure the storage array does not balloon the capacity due to inline deduplication throttling is for users to throttle the write rate on the host side low enough such that they would not exceed the flush rate of the write buffer cache. A potential downside is this assumes no inline compression savings and could add significant time to the migration. In addition, numerous migration software tools and methods currently available lack an easy way to manually throttle the bandwidth on the host side, thereby making this recommendation potentially difficult to implement.

Proposed Solution

A potential solution is to create the ability to set a "Migration QoS" option on a per volume basis. This new type of "Migration QoS" would not necessarily be optimized for latency like QoS is known today, but rather optimized by inline deduplication.

- This implementation may be specifically helpful in situations where the user may not care about the latency or ingress speed, but primarily the data reduction achieved.

- This solution could remove the complexity for the user needing to manage the migration bandwidth on the host side.

- This idea could allow for throttling to only occur for volumes listed as being migrated, thereby preserving lower latency for production workloads already present on the storage array.

- This feature could be keyed off of write buffer cache fullness or inline dedup ratio and NOT latency or IO queue depth.
About the Author

Ken Kligmann loves digging in and solving complex problems and working on the cutting edge of technology along with passionate hard working people. Ken is a Service Delivery Manager at Pure Storage, a company with unheard of NPS scores, and is responsible for Pure’s customer experience in West US region.