Openly Addressable Device Tiers

Hot Storage 2017 WACI Session
05/18/2017
Andy Kowles, Seagate, LLC.
andrew.kowles@seagate.com

The lack of determinism and the opaque nature of the existing Media Cache (MC) based Drive Managed Shingled Magnetic Recording (DM-SMR) and Host Aware (HA-SMR) designs have proven fatal for broad acceptance of shingled disks in enterprise storage systems. While the improved density and relatively high quality of SMR is welcome, systems designers cannot accept the latency variances and modal slowdowns inherent to SMR-related autonomous behavior, such as cache cleaning. This proposal aims to publicly fix that by exposing the MC (or other internal device tiers) via a simple block addressing scheme for system software to utilize. Tail latencies and overall performance for data the system designers designate can also be improved by creative use of the Openly Addressed Tiers (OATs) scheme presented.

The semantics which are possible with such a scheme are wide open and varied, and can be applied to all block storage devices (including HA-SMR) with no changes to T10 or T13 Standards. Media Cache or other tiers are explicitly addressable using the 'Shadow LBAs'.

Writes addressed to Shadow LBAs (see below) are sent to Media Cache and whether or not they are cleaned autonomously, upon command, or never could be preselected by extending the MSb addressing scheme. Reads may be done naively by the host software: The device can and will manage fragmentation and service such reads with full data integrity. Or, the host software can maintain a map of whatever data was addressed specially and avoid issuing fragmented reads IOs.

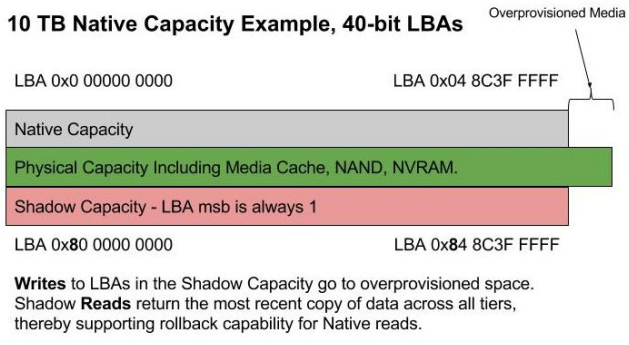

Figure 1: Shadow Addressing and Capacity

Using the simple shadow addressing, device vendors expose their investments in nonvolatile media tiers which might include a Disk Cache or NVM/SCM. Competitive advantage can be gained by device consumers by accessing these device tiers in more efficient ways specific to their unique operational profiles. In Figure 2, a batch of metadata updates can be send to the disk cache. After the velocity of metadata updates slows, the most recent copy is addressed to the Main Store (MS).

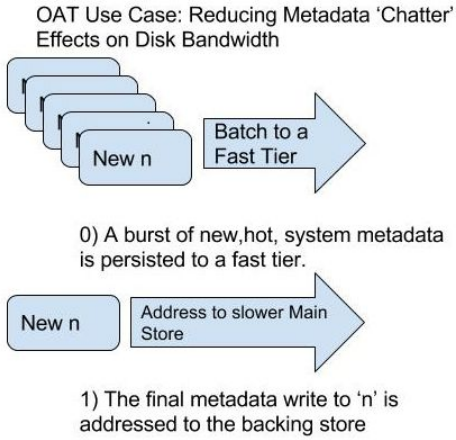

Figure 2: Using OATs to Reduce System Metadata Effects on HDD bandwidth

In this use case example, a system NAND tier is less necessary because the HDDs can absorb and service metadata IOs faster, with less total disk bandwidth and work amplification. Also note that the MC is not subject to durability issues. The final write(s) to MS cause the MC space to be freed and available for reuse.

Significantly, in another use case example, file systems (or other software) gain rollback capability of one revision for any block which is addressed to Media Cache. Also significantly, the overprovisioned disk media may be located at the OD of the disk for maximum data rate and minimum access time, or at any other radius, for example in the highest reliability area. Can low level storage software take advantage of a small but fast and reliable area of the disk or a collection of disks? Can an application be loaded more quickly from disk if it's pre-cached in MC?

Many more details and options for such open addressing, including several more use cases for DM-SMR and HA-SMR will be explored during the presentation and in subsequent work.