OPEN ADDRESSABLE DEVICE TIERS

Andy Kowles, Seagate Design Center, Longmont, Colorado | July 2017 | HotStorage

andrew.kowles@seagate.com

WHY?

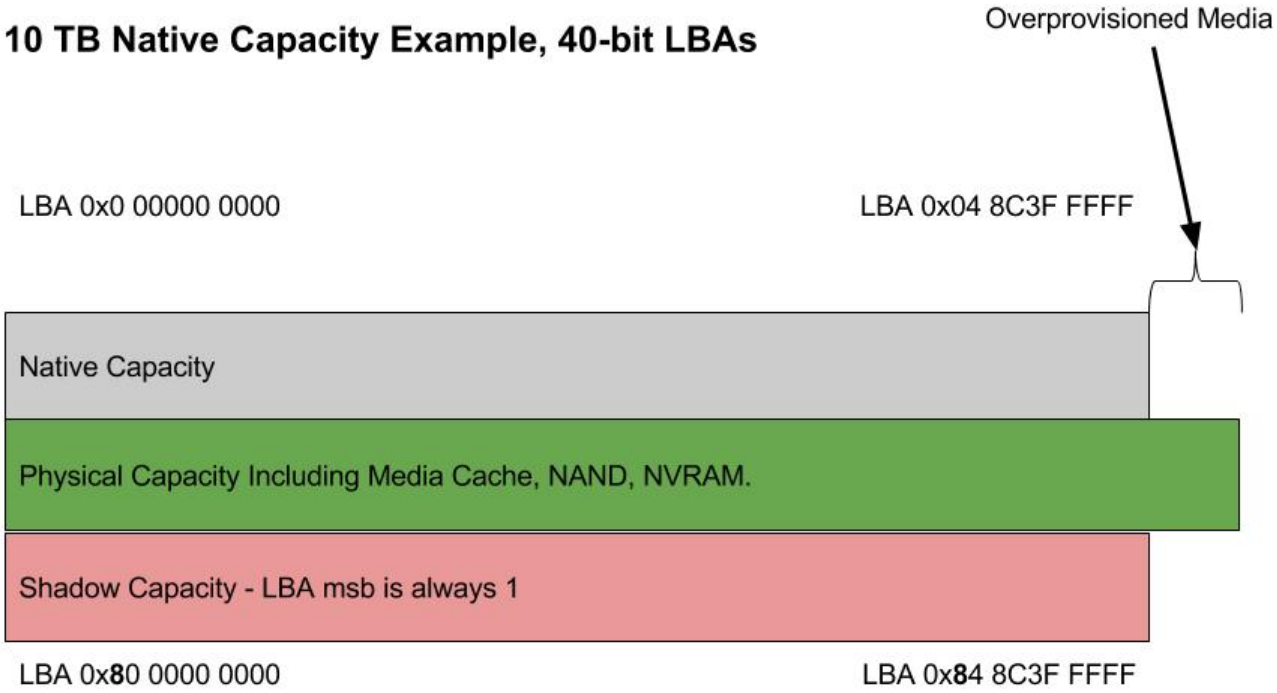
The lack of determinism and the opaque nature of the existing Drive Managed Shingled Magnetic Recording designs have proven fatal for broad acceptance of shingled disks in enterprise storage systems. The "Skylight" paper (FAST '15) is recommended reading.

"Stop doing things behind our back"

In response to this, Zone Block Device schema such as Host Managed have emerged. Such highly restrictive access models improve the drive latency responses and predictability, versus Drive Managed.

A complementary, wild and crazy way to address HDDs is offered here: Using LBA addressing tricks and explicit policies, *allow storage software developers to access internal device tiers* which have been, until now, hidden and autonomous. No T10/T13 changes are strictly required.

WHAT IS IT? PART 1: "SHADOW" ADDRESSING



Writes to LBAs in the Shadow Capacity go to overprovisioned space.
Shadow **Reads** return the most recent copy of data across all tiers, thereby supporting rollback capability for Native reads.

This addressing scheme is but one option for direct addressing of device tiers.

WHAT? POLICIES

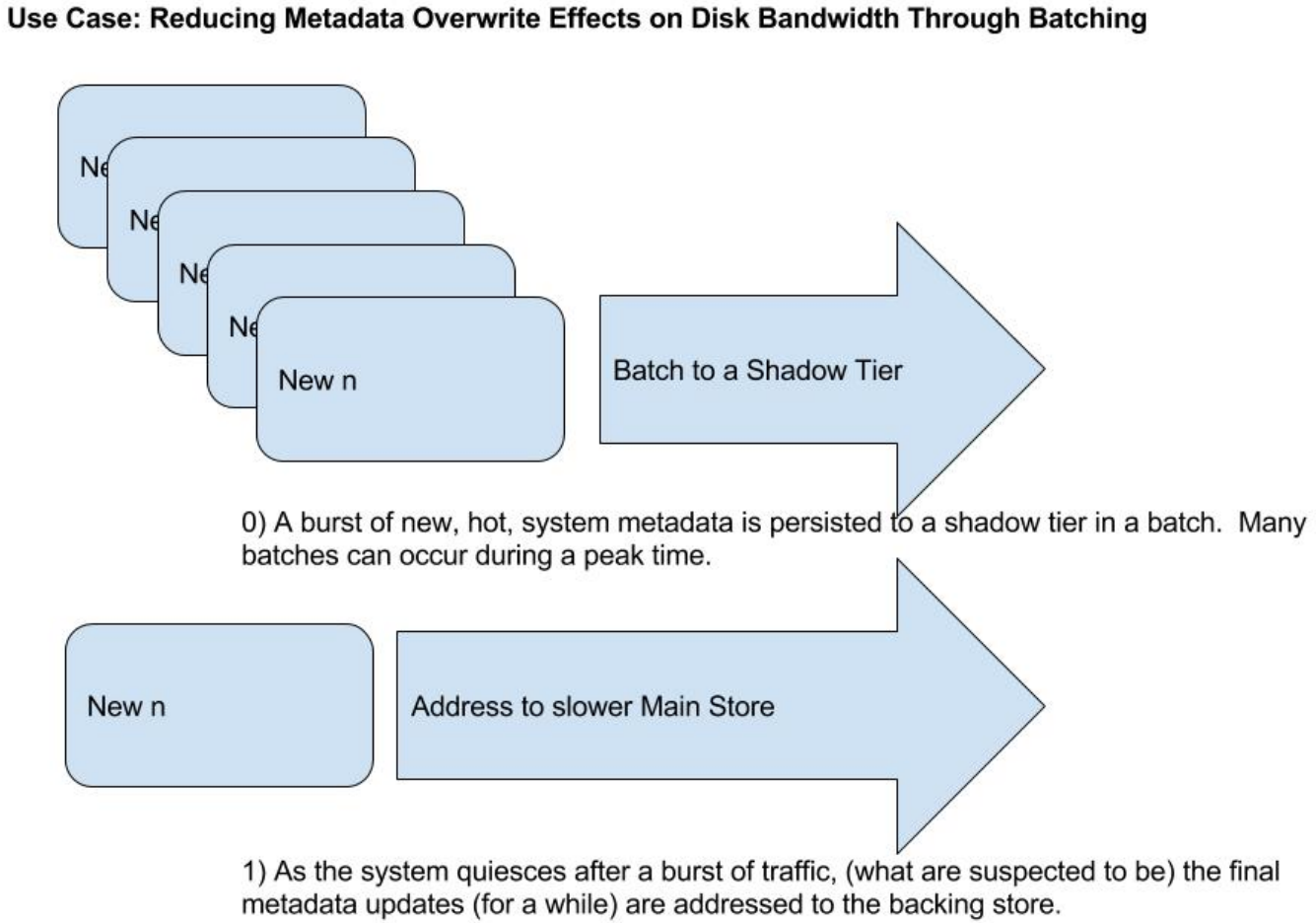
A “tier” on a device defined similarly as one in a system: It is a storage area of a particular size, cost, performance, availability, and reliability.

Policies for OATS are undiscovered country. What policies would you prefer to see? Here are some possibilities…

- Latency/Performance: Cleaning from the tier NEVER occurs. Strict LRU replacement policies inside the tier.
- Latency/Performance: Cleaning from the tier NEVER occurs. Once Shadow address space or tier is full, Shadow Addressing has no effect.
- Reliability: Data is Duplicated when directed to Shadow.
- ACID write transactions without Flush Cache (fsync) or Disable Write Cache, or FUA, are performed on data directed to Shadow addresses.
- With multiple tiers, different data can be directed to tiers with the appropriate reliability and performance properties.

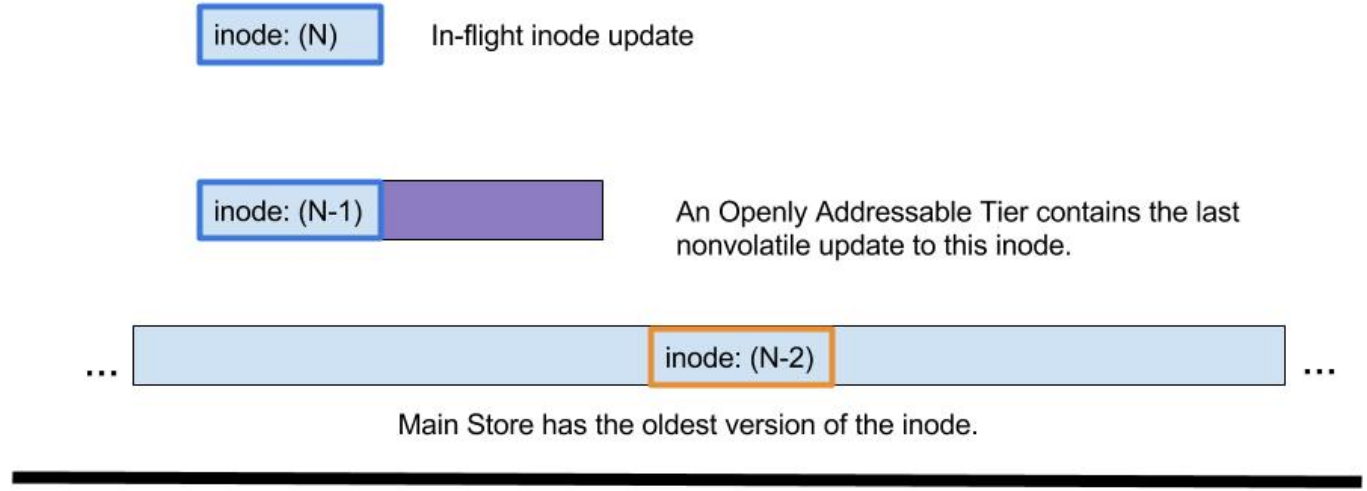
EXAMPLE: BATCHED METADATA UPDATES

Analogous to the way ext4-lazy (FAST'17) improved SMR disk performance by using a large journal and avoiding excessive in-place inode updates.



Could this reduce the required size of NAND layers above the HDD in a data center?

Use Case: Rollback, Write Transactions, and a Reduced Need for fsync/cache flush



The system, then, can roll back a partial update at time N (for example) in the face of exceptions or uncontrolled loss of power.

System Rollback: A Trim/Unmap command using the Shadow addresses, followed by a read, would return Version (N-1) across the device.

QUESTIONS AND FURTHER STUDY

- What are the most useful use cases to storage software developers?
- How many upper LBA bits would be used, and for what? (48bits of LBA addressing for 4KiB blocks provides for 1125 PB of capacity. At 512B blocks, 140 PB)
  - 1 bit for Shadow A → 70PB
  - 1 bit for Policy X or Alternate Tier B → 35PB
  - 1 bit for Duplication/Reliability Settings.
- Can Host Managed or Host Aware Zone Block Devices be improved?
- Further Study: Implement an openly addressable device using an existing DM-SMR or SSHD device. Incorporate "Shadow Policies" derived from use cases into the device. Effect changes to the block stack to harness the Openly Addressable Tier for the desired use case(s). Test and measure and report.
- Very interested to hear you ideas, in person or otherwise.

Seagate Global Firmware
andrew.kowles@seagate.com | 720.684.8469