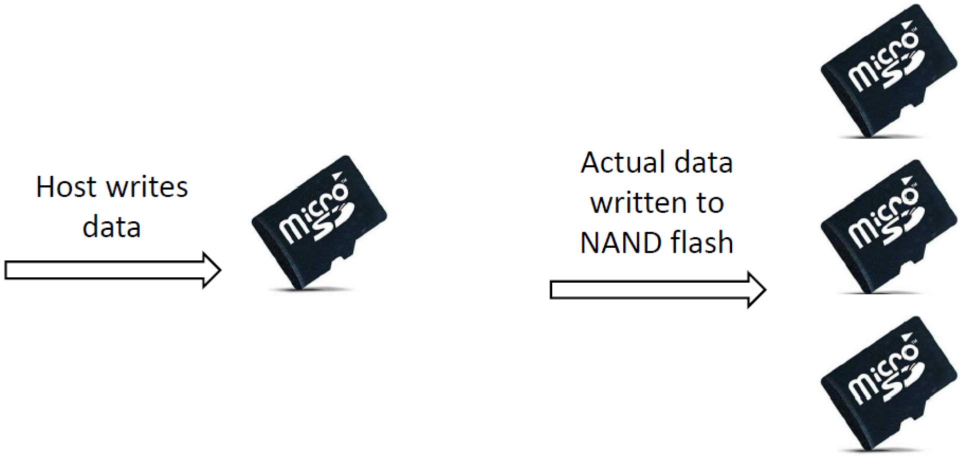
How to reduce flash storage system write amplification

weimingchang@huawei.com

What is system write amplification

What is WAF?

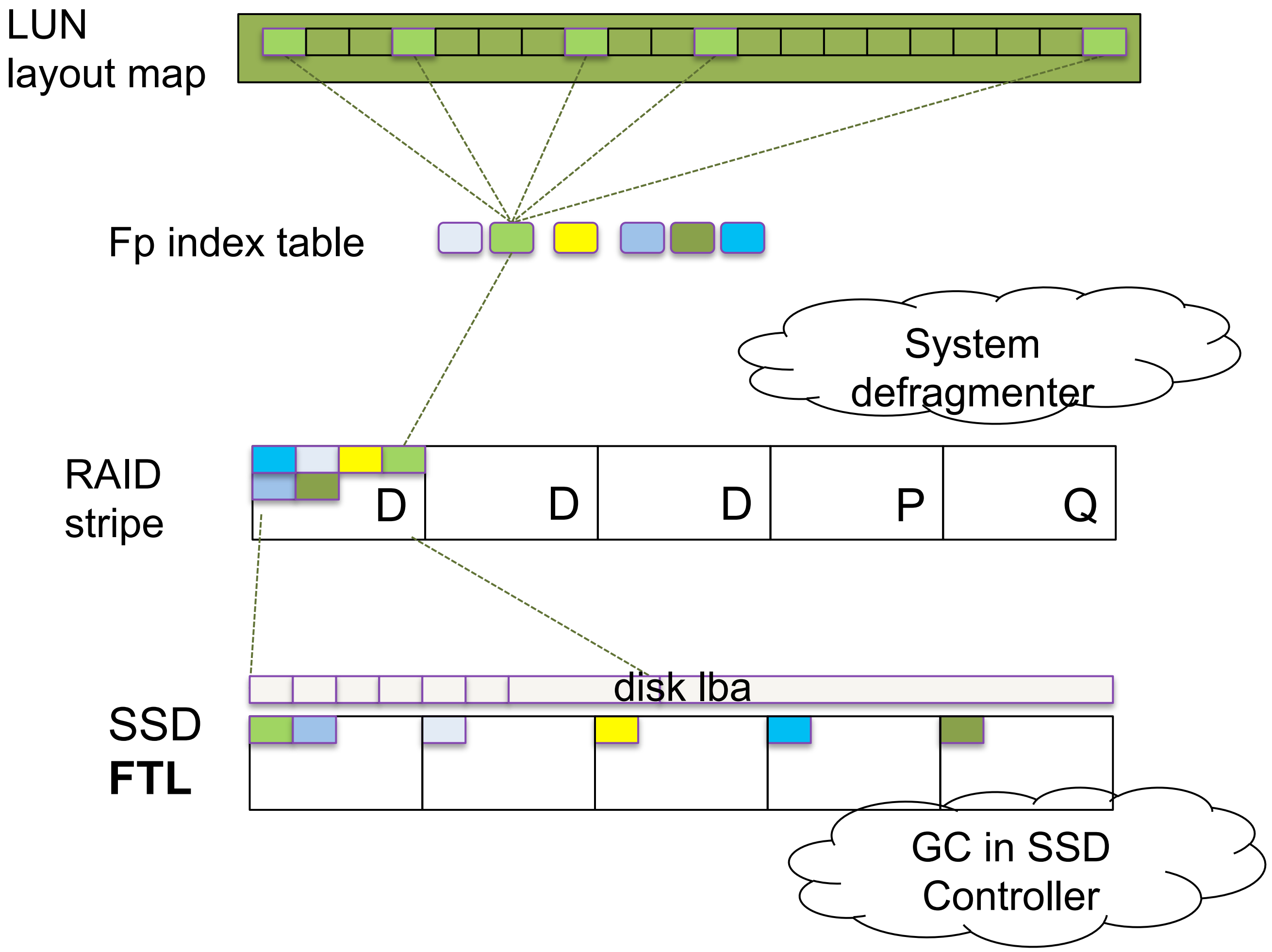
$$\textit{Write Amplification Factor (WAF)} = \frac{\textit{Data written to NAND Flash}}{\textit{Data written by host}}$$



Host writes data

Actual data written to NAND flash

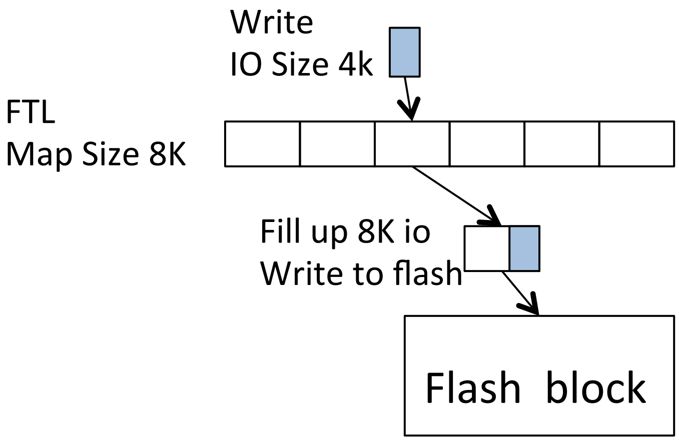
Flash storage system



IO write amplifier

Systems have different mapping tables and define different block size

- Remote Copy block size
- Snapshot block size
- Dedup block size
- compress block size
- FTL page size of SSD



the system needs according to various scenarios IO model, unified consider defining each block size. Match io model and Reduce write IO amplification.

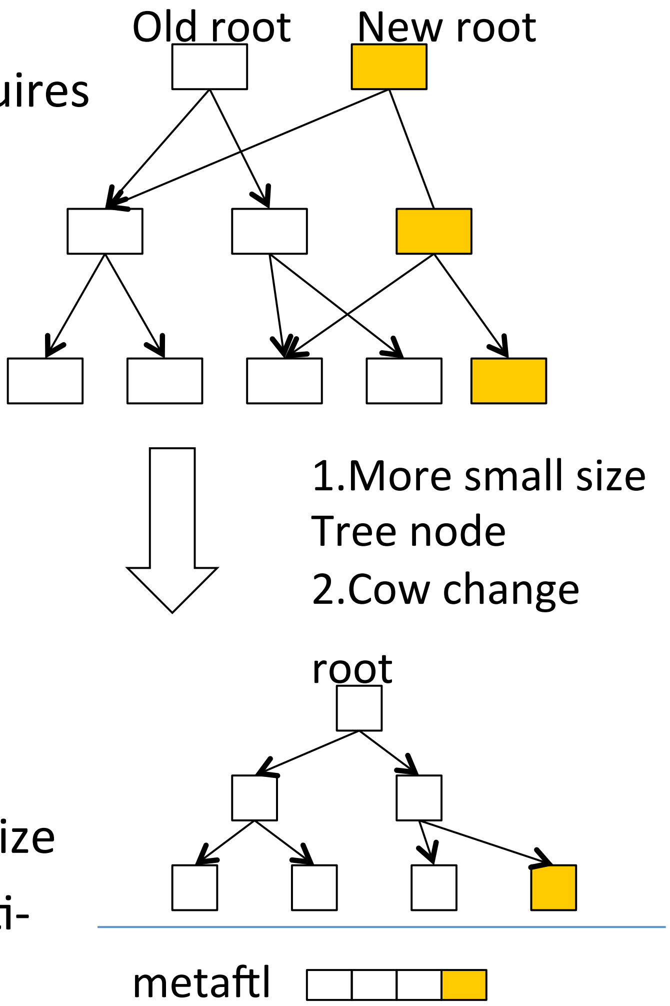
Meta data write amplifier

There are various kinds of metadata system requires maintenance

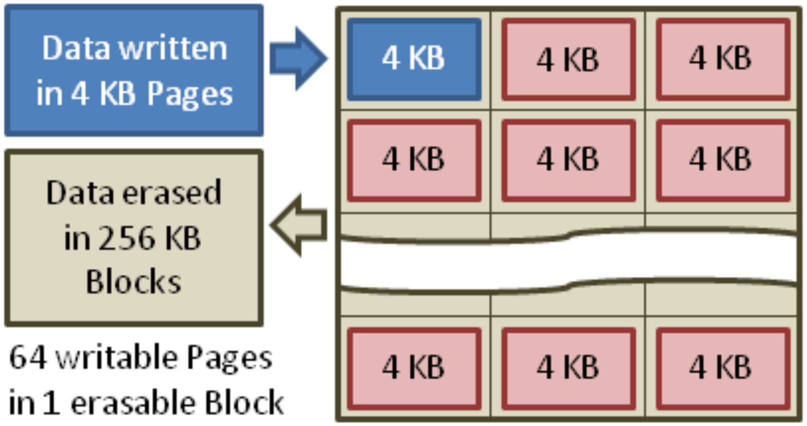
- Storage Space alloc and free infomation
- Dedupe fingerprint table
- Data referen count table
- LUN or file layout infomation
- SSD FTL table

To cutdown metadata write amplification.

- More smaller metadata management block size
- Add meta ftl, ROW metadata, and avoid multi-level modify metadata
- Cache hot metadata on NVRAM



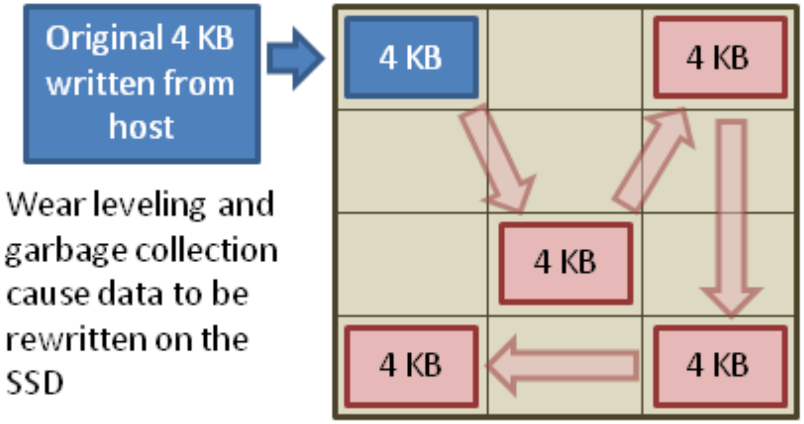
garbage collection write amplifier


Data written in 4 KB Pages

Data erased in 256 KB Blocks

64 writable Pages in 1 erasable Block

Typical NAND Flash Pages and Blocks

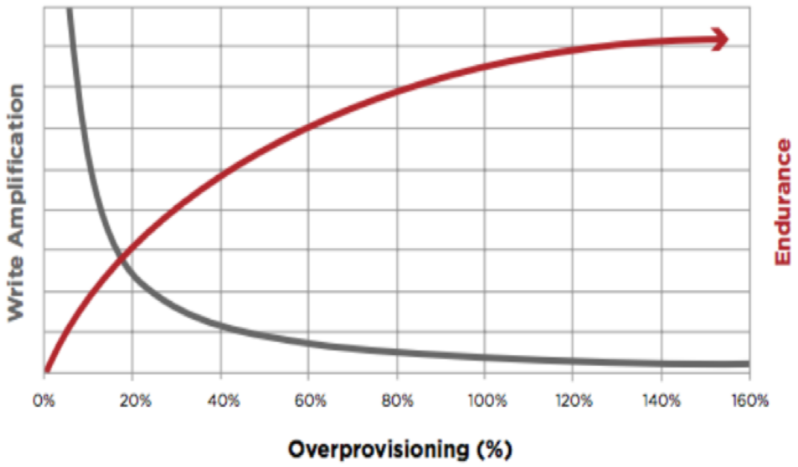

Original 4 KB written from host

Wear leveling and garbage collection cause data to be rewritten on the SSD

Solid-state drive Flash memory

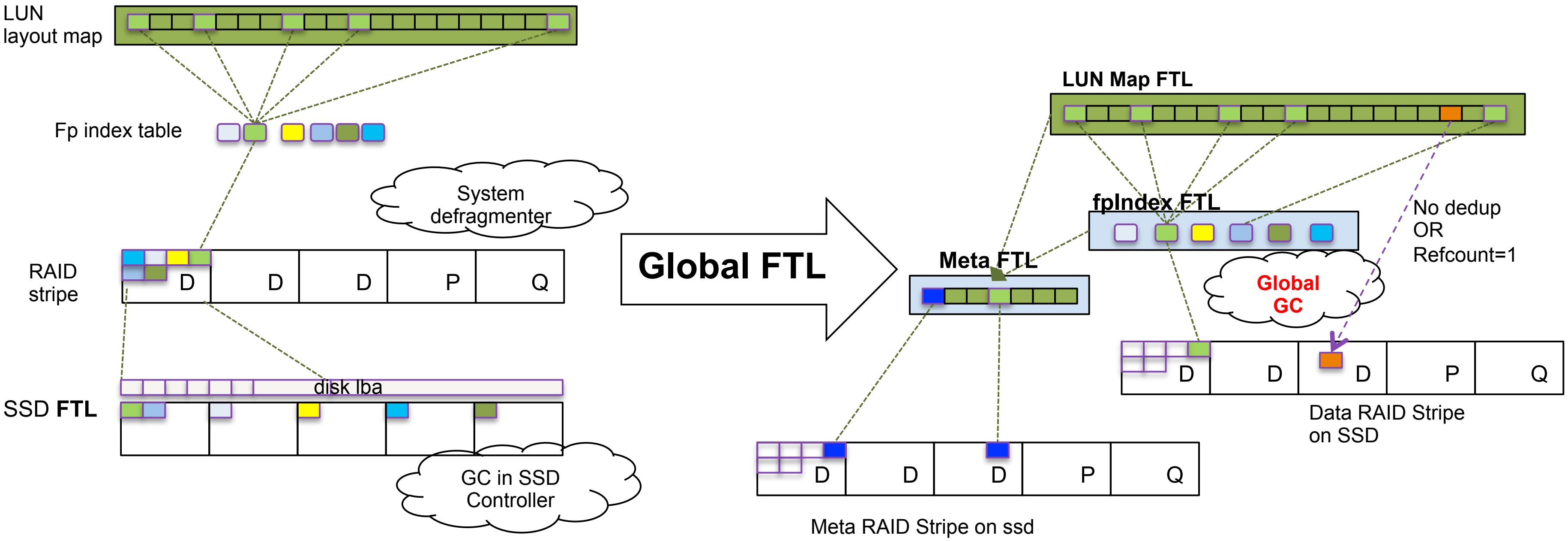
GC write amplifier decide by OP .
More bigger OP can get more smaller GC write amplifier.

Figure 2: Write Amplification vs. Over Provisioning

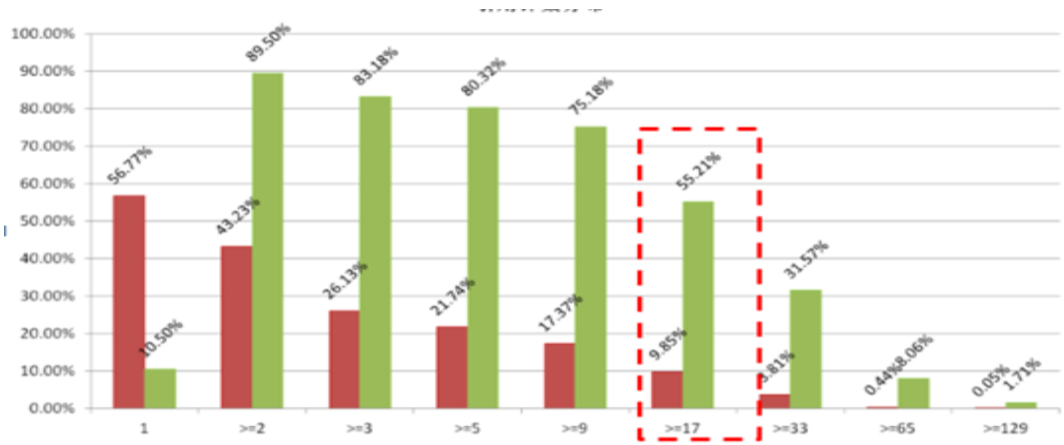


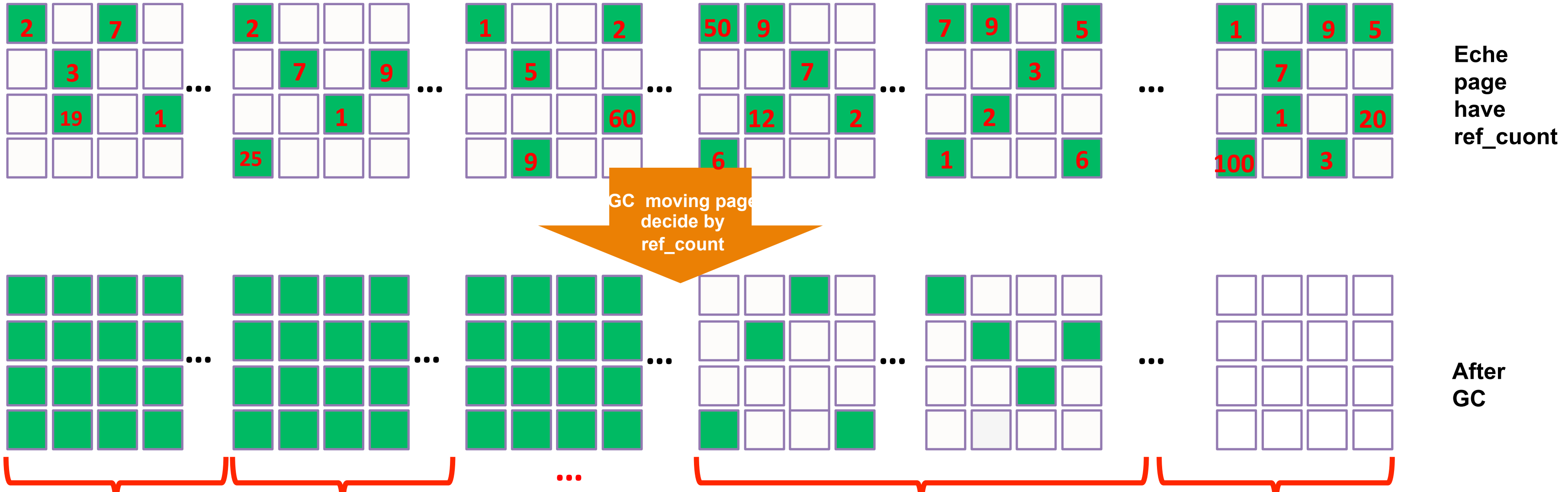
Global Dedupe and GFTL for GC

- System supports global dedupe and GFTL, All users unused capacity as OP, recycle the invalid blocks space, erase invalid data blocks, the reconstruction only valid data block.



Separating Static and Dynamic Data





Eche page have ref_cuont

After GC

RAID Stripe for Ref_count>5 page

RAID Stripe for Ref_count[2~5] page

…

RAID strip for Hot page

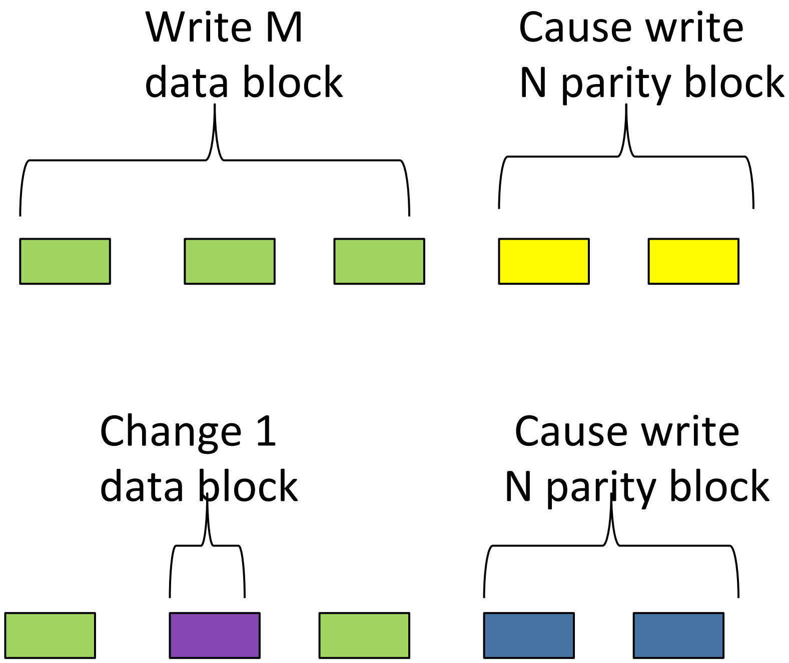
Idle Block

RAID parity and log write amplifier

- M write data to calculate the N parity protection, cause additional parity data write amplifier

- To avoid overwriting the write process error when system reset, resulting in inconsistent stripe, system need to first write the stripe change log, ensure system can restore consistency strips, such cause a change log write amplification

by garbage collection ,New write data always write new RAID stripe reduce parity block writing amplifier.



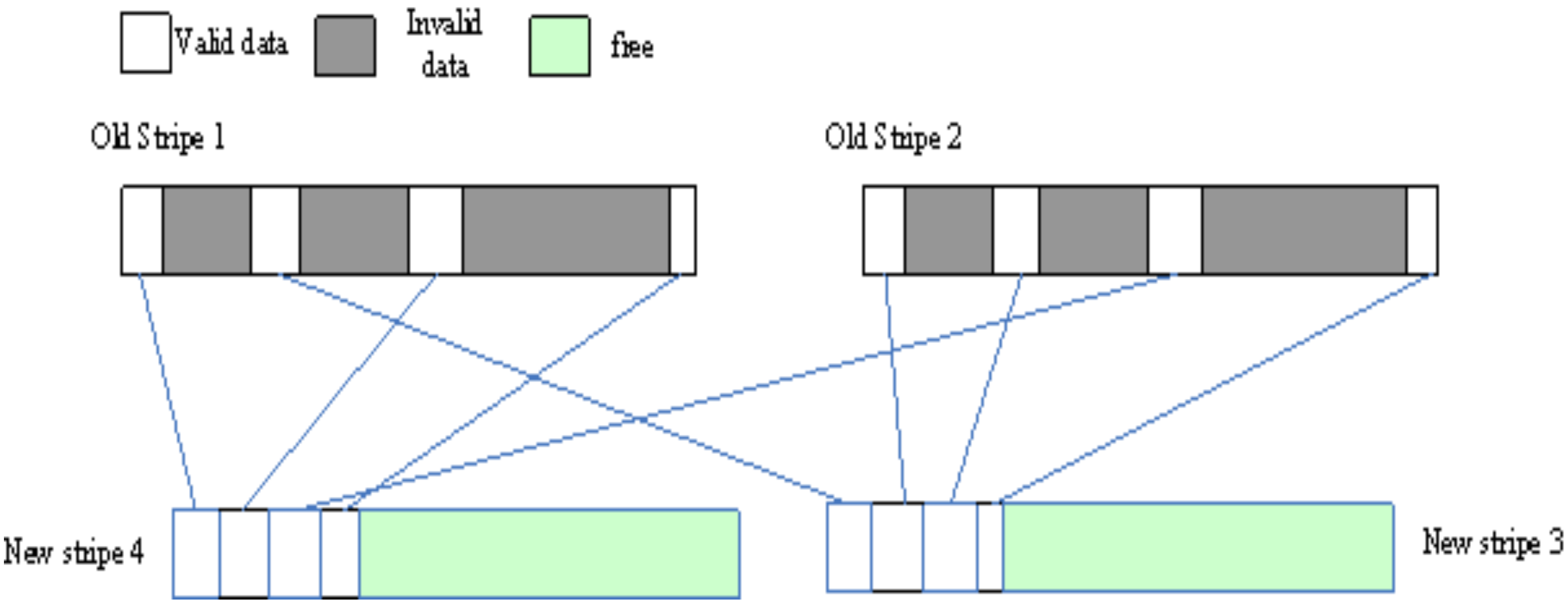
Write M data block

Cause write N parity block

Change 1 data block

Cause write N parity block

Flash aware RAID reduce Rebuild WA



Based on the data validity, reclaim and reconstruct stripe, reduce rebuild stripe numbers (old stripe can free directly)

conclusion

- set the appropriate block size for IO model of different scenarios, different LUN can have different block size. Avoid not alighed IO write amplification

- More smaller metadata block size, and add a metaftl to avoid modify multi-level tree metadata, Cache all hot metadata on NVRAM can cutdown metadata write amplification .

- the system can uses the global garbage collection and avoiding the SSD inside each garbage collection. To reduce parity block write amplification.

- Systems do global garbage collection By separating Static and Dynamic Data. in the same system OP size, increasing the number of garbage data blocks in the hot space, will reduced GC write amplification.

Thank You