Getting the Most Out of SSD: Sometimes, Less is More

Bruce Moxon

Chief Solutions Architect

STEC

Overview

- A Quick Solid State Backgrounder
  - NAND Flash, Log structured file systems, GC, and Overprovisioning – less really is more!
- Benchmarking
  - Operationally representative testing
- Applications
  - Why Flash, why now?
  - Caching – Less *is* more Enough
  - Optimizing the Stack and Changing Application Architectures

Solid State Performance Characteristics (General)

HDD (SAS)	Sequential	Random
Read	200 MB/s	200 IOPS
Write	200 MB/s	200 IOPS

2-8x ↓ ↓ 100-1000x

SSD / PCIe	Sequential	4K Random
Read	.5 – 1.5 GB/s	60-200K IOPS
Write	.3 – 1 GB/s	15-40K IOPS

Rand Response	Read	Write
HDD	8 ms	0.5 ms*
SSD	60 us	20 us

General Performance Characteristics. YMMV, depending on

- Device architecture
- Interface (SATA, SAS, PCIe)
  - 2-4x performance range
- Transfer sizes
- QDs (concurrency)

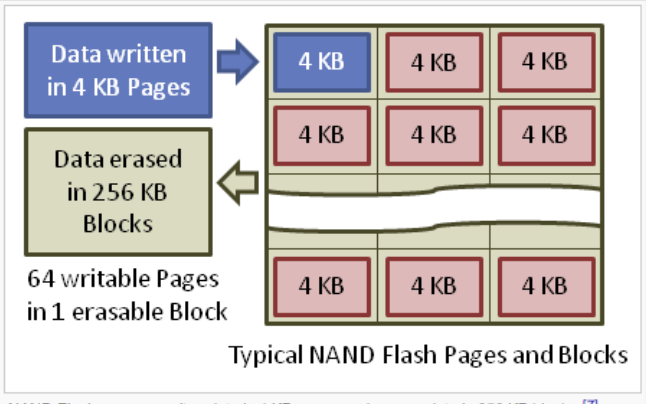
Cost Differential

- $0.50 - $1.50 / GB SAS HDD
- $2 - $12 / GB SSD/PCIe

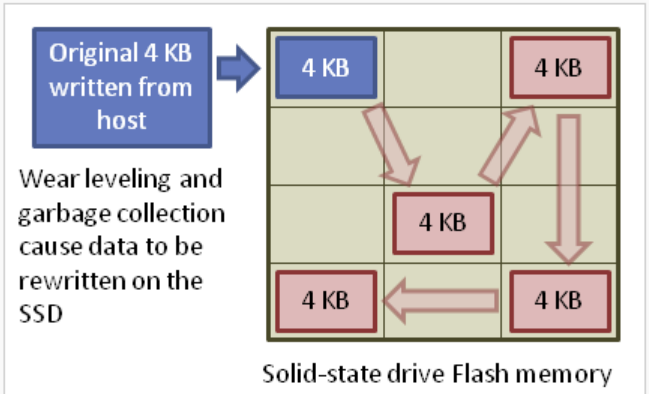
Sweet Spot

- High Random IOPS (esp. Read)
- Low Latency

Solid State Storage Fundamentals



NAND Flash memory writes data in 4 KB pages and erases data in 256 KB blocks.[7]



An SSD experiences write amplification as a result of garbage collection and wear leveling, thereby increasing writes on the drive and reducing its life.[1]

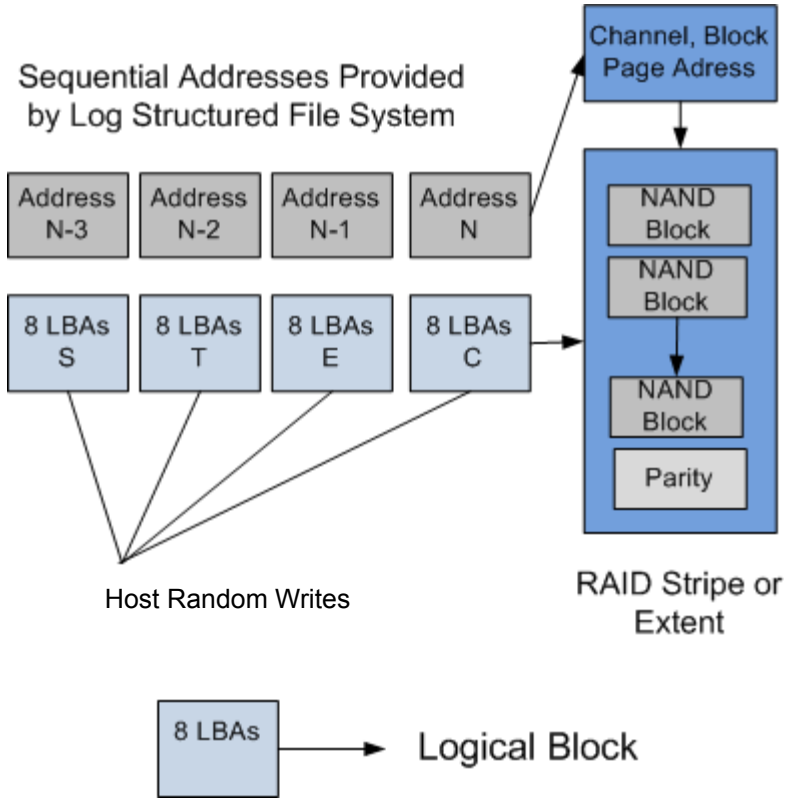
Everything I Needed to Know I learned at FMS …

- Data is read/written in pages (typically 4-8 KB)
- Data is *erased* in multi-page blocks (e.g., 128)
- Data can only be written (programmed) into a previously erased block (no “overwrite”)
- Background garbage collection (GC) copies valid pages to “squeeze out” deleted pages and make room for new data
  - OS/FS integration (TRIM)
- Additional *write amplification* can occur in support of wear leveling and GC
- Flash memory can only be programmed/erased (P/E) a limited number of times (Endurance)
  - *Performance degrades over device lifetime*
- Overprovisioning is usually employed to afford more flexibility in background tasks such as wear leveling and garbage collection. It reduces write amplification, in turn extending the life of the device
  - Background tasks affect performance, especially at “peak” device speeds (latency)
- Implementations vary widely

http://en.wikipedia.org/wiki/Write_amplification#Basic_SSD_operation

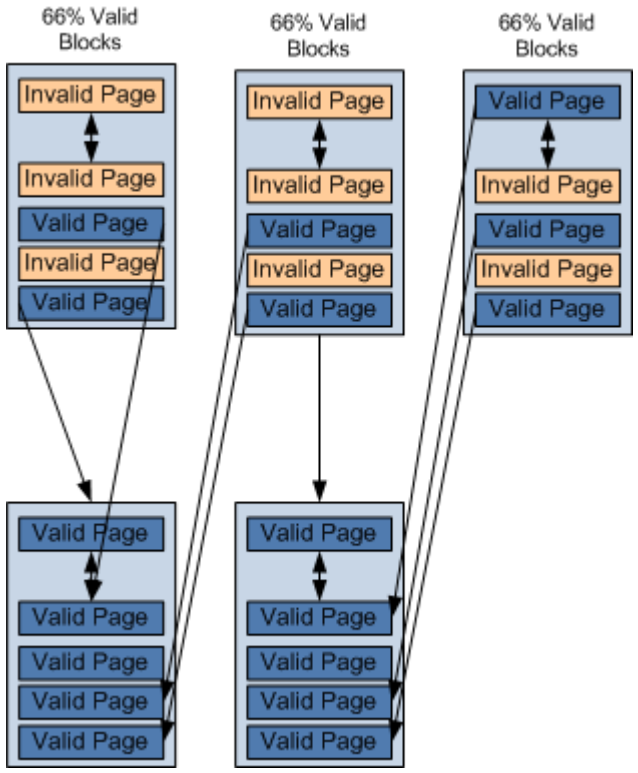
Log Structured File System

- First Described in a paper by
  - Mendel Rosenblum and John K. Ousterhout
    - http://www.cs.berkeley.edu/~brewer/cs262/LFS.pdf
  - Basic principles
    - Extra space or over-provisioning is added to the capacity of the drive
    - Transforms random writes into sequential writes
      - The percentage of transformation is dependant on the amount of over-provisioning
    - The physical location of write data is allocated on writes
      - The location of a LBA is never in the same physical location on back to back writes to media



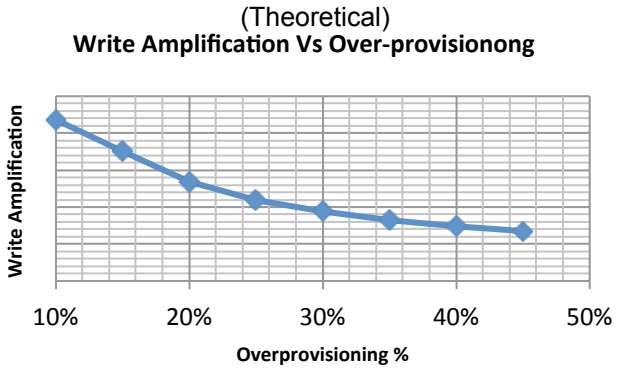
Garbage Collection Example: Random Writes

- In this example the write amplification is equal to 3
  - 3 Blocks have 66.66% invalid Blocks
  - 3 * 66.66% = 2 Full Blocks
- The valid data is read from 3 blocks that are 66.66% full and written to 2 new blocks
  - The 2 new blocks now contain 100% valid data
  - The old three blocks are then erased
- To free up one block worth of data requires two the equivalent of two blocks of data movement
- The write amplification is equal to 1 + the number of blocks moved = 3



Performance vs. Formatted Capacity

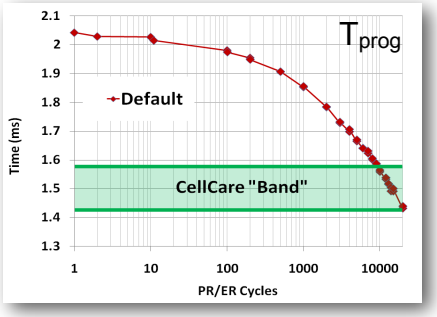
- Write Amplification increases as the formatted capacity increases
  - Higher WA lowers random write performance and PBW
- Low WA and high PBW is a requirement in heavy-write, enterprise class storage applications
- High WA reduces Endurance
  - Advances “end-of-life” device degradation

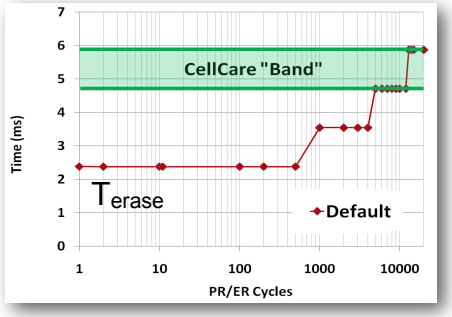


Theoretical Device Performance		
Capacity (GB)	Seq write (MBs)	8K Random Writes (IOPS)
1000	500	12.0
980	500	14.0
960	500	16.3
900	500	20.7
850	500	24.3
800	500	27.6

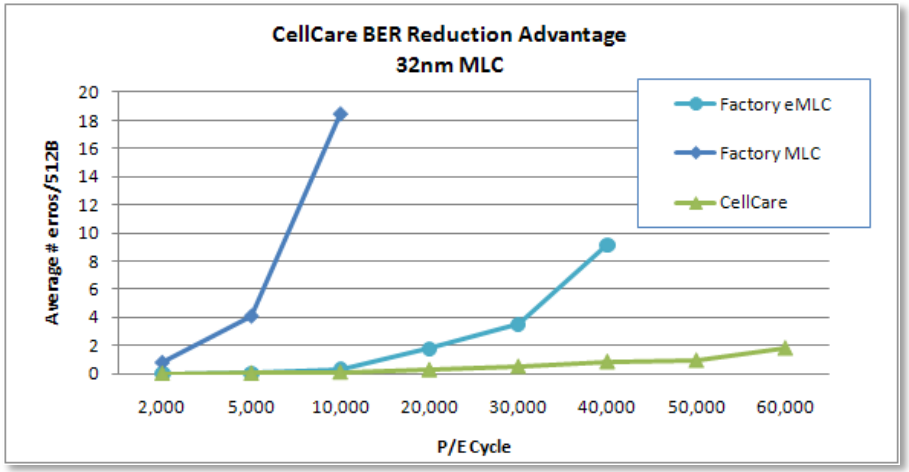
Flash Endurance

- Number of Program/Erase (PE) cycles on Flash is Limited
- Performance Characteristics of the device changes over time





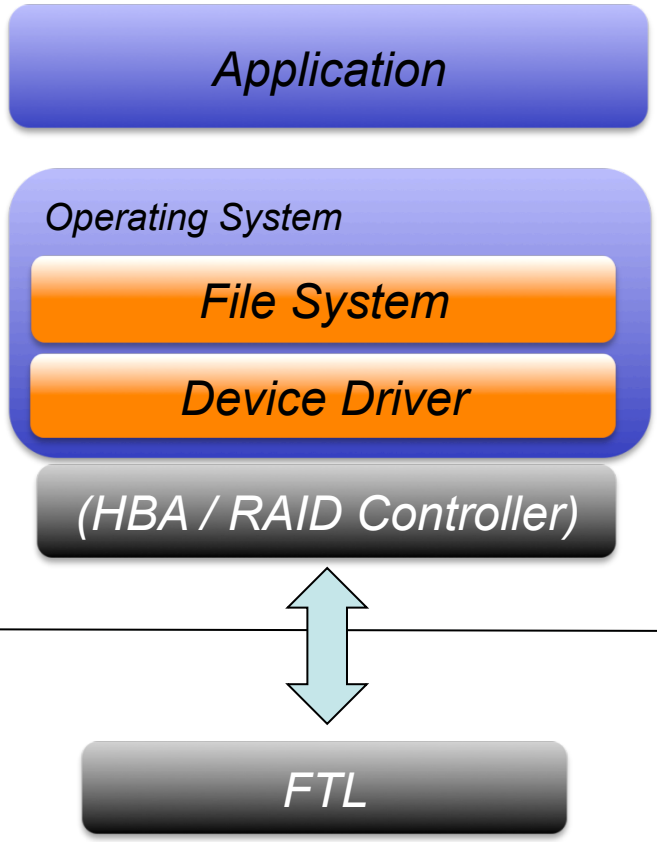
- STEC's Cell Care:
  - Dynamically optimizes read, write, and erase levels over the life of the drive
  - Employs DSP-based error correction
- To Deliver:
  - *Increased Endurance*
  - *Improved Reliability*
  - *Consistent Performance*



Lies, damn lies, and Benchmarks …

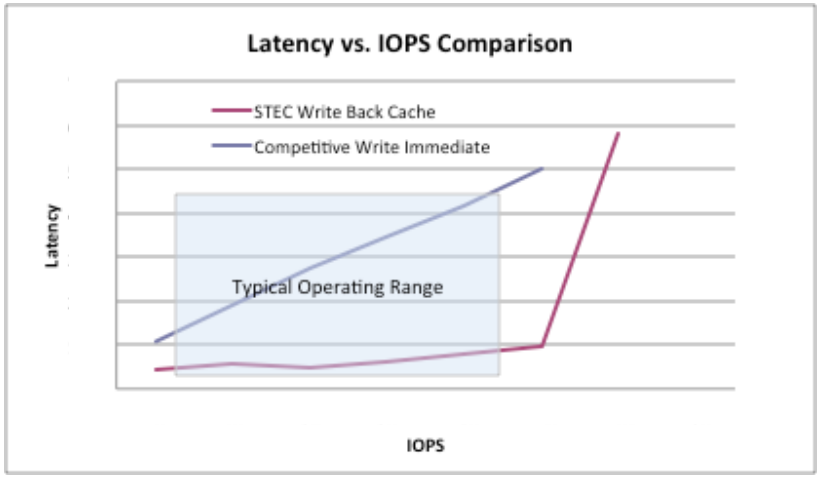
- But Seriously … (Mostly) well-intentioned, sometimes misguided
- But what are we benchmarking?
  - Overprovisioning
  - Endurance and Consistent Performance considerations
  - Full or near-full devices (preconditioning)
  - Compression and de-duplication
- And more importantly …
  - How does that relate to the expected Operational Environment? (vs. “4-corners”)

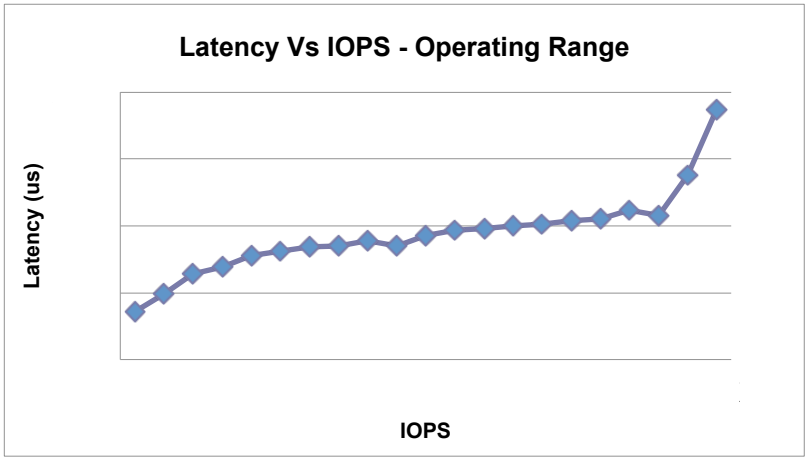
The Big Picture



Design .. and Benchmark .. for Operations

- You *never* design to the edge of the engineering envelope
  - But you probably do want to understand where the edge is
  - Consolidation scenarios
- Lots of “4-corners” device benchmarks
  - rand/seq read/write
- Few device-level benchmarking tools support throttling
- Multi-disk JBOD/RAID
- Application benchmarks tweaked to show off SSD
  - Low memory DB configs





Summary

- Understand the relationship between overprovisioning, endurance, and performance in order to make appropriate trade-offs
- Single device benchmarks have their place, but may not be all that useful in predicting application performance
  - Strive to benchmark operationally relevant scenarios