Overcoming Reductions in NAND Endurance Ratings

JB Baker
Sr Director Product Management

ScaleFlux
Framing the Endurance Challenge

Innovations in SSD Endurance beyond LDPC
  • Transparent / Drive Integration
  • Storage Driver Integration
  • Application Integration
  • Future
NAND & SSDs: Better, Faster, Cheaper?

Cheaper
- Enterprise SSD $/GB
  - 1/10th the price and falling

Faster
- SSD Throughput (MB/s)
  - 10x throughput with 20x soon

Better
- NAND Endurance (k P/E)
  - 1/20th the life @ TLC
  - 1/100th the life @ QLC

*Source: Forward Insights, SSD Insights Q1’19
The Endurance Challenge

Total Bytes Written (TBW) = Raw Capacity * Program-Erase Cycles

Write Amplification

NAND endurance decline outpaces Capacity Growth

Need Write Amplification Innovations to Contribute!

*Source: Forward Insights, SSD Insights Q1'19
Agenda

- Framing the Endurance Challenge
- Innovations in SSD Endurance beyond LDPC
  - Transparent / Drive Integration
  - Storage Driver Integration
  - Application Integration
  - Future
**In-line Compression/Decompression**

**Transparent / Drive Integration**

- **What it is:**
  - Encodes the data to reduce the physical space it consumes
  - Runs a compression algorithm on data as it is written to Flash
  - Decompresses data upon read

- **What benefits it can deliver**
  - Increased *effective* overprovisioning (OP)
  - Significant reduction in Write Amp → Increased TBW
  - Improved IOPs and Latency → Reads & Writes
  - Additional User Space

- **Limitations / requirements to derive the benefit**
  - Data compressibility varies… but a little goes a long way!!
In-line Compression/Decompression
Transparent / Drive Integration

<table>
<thead>
<tr>
<th>Raw Capacity</th>
<th>User Capacity</th>
<th>Rated OP %</th>
<th>Effective OP% with compression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.2:1</td>
</tr>
<tr>
<td>4TB</td>
<td>3.2TB</td>
<td>28%</td>
<td>54%</td>
</tr>
<tr>
<td>3.84TB</td>
<td>3.2TB</td>
<td>7%</td>
<td>60%</td>
</tr>
</tbody>
</table>

- Minimal Compression lets “7% OP” act like “28% OP” → Similar endurance & performance with 20% more user space!
- Moderate Compression yields ~100% Effective OP → Enables Write Amp close to 1... doubling or more the TBW!!

Data Compressibility Examples:
- <1.2:1 – Images, Video, Encrypted
- 1.2:1 – Binaries, DLL, EXE
- 2:1 – XML
- >2:1 – HTML, Logs, Database

See Thomas McCormick’s preso from FMS 2016 for detailed WA vs OP: [link](https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160809_FC12_%20McCormick.pdf)
Atomic Write

Storage Driver Integration

What it is:
- Atomic write operations guarantee that either “all specified blocks are written” or “no blocks are written”

What benefits it can deliver:
- Turn off double-write buffer (DWB) for databases
- Cut writes to NAND by 50% → 2x SSD Endurance
- Cut writes per transaction by 50% → 2x QPS*

Limitations / requirements to derive the benefit:
- Filesystem must guarantee that write requests occupy consecutive LBAs
  - E.g. EXT4/bigalloc used so that MySQL/InnoDB data unit is in one 16kB page

*SysBench Write-Only benchmarks
Atomic Write
Storage Driver Integration

- Database Table update = 2 Writes
  - 1st write to DWB
  - 2nd write to Database Table
  - Good data is recovered from DWB in case of crash

- Database Table update = 1 Write
  - Data guaranteed to be updated completely, or not at all
  - 50% Less data written to NAND
  - 50% Fewer write I/Os
Streams
Storage Driver Integration

What it is:
- The Streams Directive enables the host to indicate (i.e., by using the stream identifier) to the controller that the specified logical blocks in a write command are part of one group of associated data. This information may be used by the controller to store related data in associated locations or for other performance enhancements.*

What benefits it can deliver
- Performance & Endurance improvements
  - Separates Read/Write queues ✓
  - Set unique OP levels for each Stream ✓
- Avoid Garbage Collection for long-term data → Reduce WA ✓
  - Manage free/erase block pools separately for each Stream

Limitations / requirements to derive the benefit
- Host awareness of the Streams
- Benefit varies widely depending on the size & update frequency of the Streams relative to each other

*NVM-Express™ Revision 1.4, Sect 9.3
Streams
Storage Driver Integration

- **Multi-Stream**
  - Like data logically stored together
  - *Isolate GC & I/O traffic for each data type*

- **Single Stream**
  - All data jumbled together
  - No logical separation for GC, OP or Read/Write

**Database Logs**
- Small Random
- Re-written Often

**Files**
- Large Sequential
- Long Term Storage

**Analytic Data**
- Medium Sequential & Random
- Temporary Storage

**Small Random**
- Re-written Often
Group Garbage Collection
Application Integration

- **What it is**
  - Consolidation of Garbage Collection activities between the application and the SSD

- **What benefits it can deliver**
  - Eliminates redundant GC $\rightarrow$ Reduction in WA
  - Higher throughput & less latency variability
  - Zero-OP SSD $\rightarrow$ Adds 7%, 28%, or more to usable GB

- **Limitations / requirements to derive the benefit**
  - File System or Application must initiate GC, compaction or defrag
    - E.g. RocksDB, ZFS, Aerospike
  - FS or Application changes to communicate with the SSD Firmware
  - SSD Firmware capable of informing FS/App of the physical location
    - E.g. Open Channel
Group Garbage Collection
Application Integration

Baseline
Separate Defrag & GC

Av1 106 KOPS Steady State
166M 4K Records
User Capacity 1600GB

0% OP Flash Storage
Group Defrag & GC

Avg 115 KOPS Steady State (+8.5%)
215M 4K Records (+30%)
User Capacity 2048GB (+28%)

WA Reduction improves:
Endurance, Capacity, Consistency, Performance

POC Results with CSS 1000
Modifications to Aerospike group defragment

448GB OP
1600GB User

2048 GB
Physical Flash

2048 GB
Physical Flash

2048GB User
Future...

- **Global FTL**
  - Manage the NAND across SSDs as a single pool
  - Cut RAID overhead by 50% → single level vs Host & In-Drive
  - Global OP / wear leveling
  - Efficient support for large numbers of sets

- **Deduplication**
  - Replace multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth
  - More effective with larger data sets

- **Larger Compression Blocks**
  - Yield higher compression ratios
  - Tradeoff with Read performance
Thank You

JB Baker, Sr Dir Product Management
ScaleFlux
Jb.baker@scaleflux.com