Flash Memory Summit

Overcoming Reductions in NAND Endurance Ratings

JB Baker

Sr Director Product Management

ScaleFlux™

Agenda

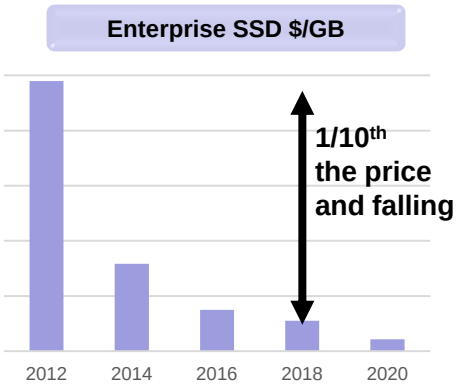
Flash Memory Summit

- Framing the Endurance Challenge
- Innovations in SSD Endurance beyond LDPC
  - Transparent / Drive Integration
  - Storage Driver Integration
  - Application Integration
  - Future

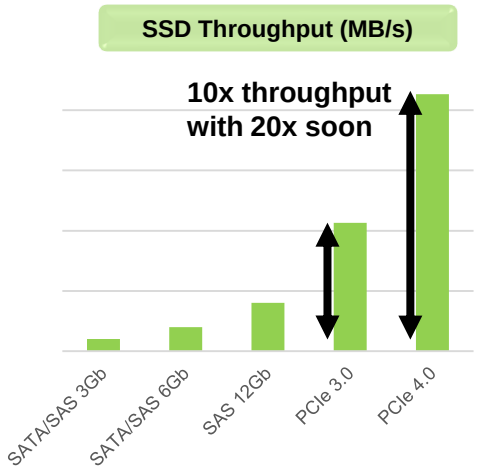
NAND & SSDs: Better, Faster, Cheaper?

Flash Memory Summit

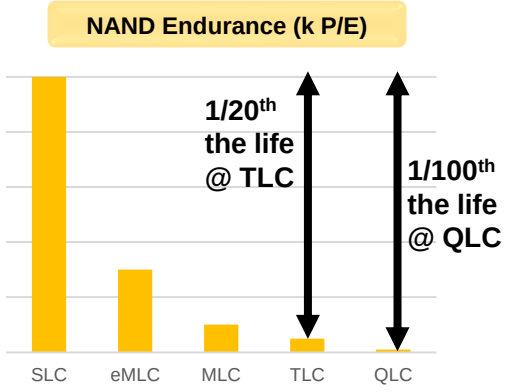
Cheaper



Faster



Better

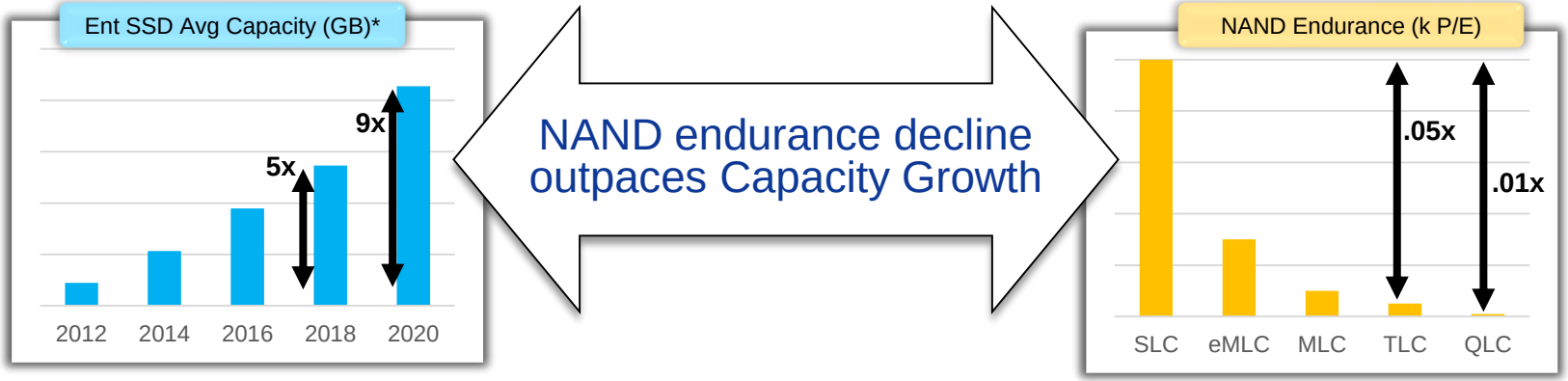


*Source: Forward Insights, SSD Insights Q1'19

The Endurance Challenge

Flash Memory Summit

$$\text{Total Bytes Written (TBW)} = \frac{\text{Raw Capacity} * \text{Program-Erase Cycles}}{\text{Write Amplification}}$$



Need Write Amplification Innovations to Contribute!

*Source: Forward Insights, SSD Insights Q1'19

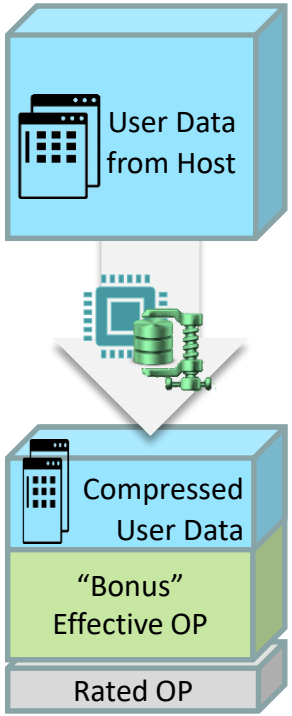
Agenda

- 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

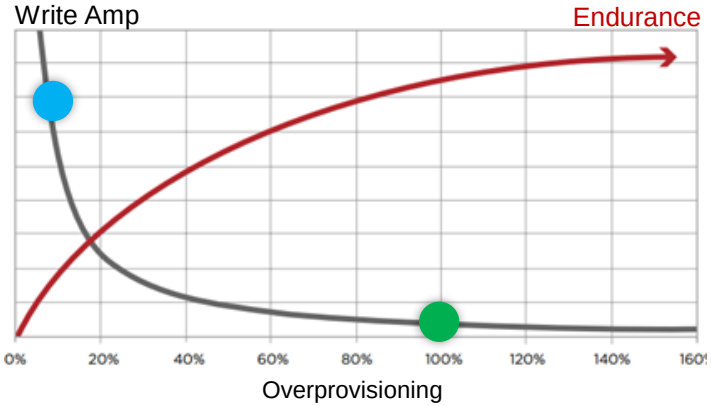
- What it is:
  - Encodes the data to reduce the physical space it consumes
  - Runs a compression algorithm on data as it is written to Flash
  - Decompresses data upon read
- What benefits it can deliver
  - Increased *effective* overprovisioning (OP)
  - Significant reduction in Write Amp → Increased 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

Raw Capacity	User Capacity	Rated OP %	Effective OP% with compression		
			1.2:1	1.5:1	2:1
4TB	3.2TB	28%	54%	***92%***	156%
	3.84TB	7%	***28%***	60%	***113%***



***Moderate* Compression yields ~100% Effective OP**
→ Enables Write Amp close to 1… doubling or more the TBW!!

***Minimal* Compression lets "7% OP" act like "28% OP"**
→ Similar endurance & performance with 20% more user space!

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

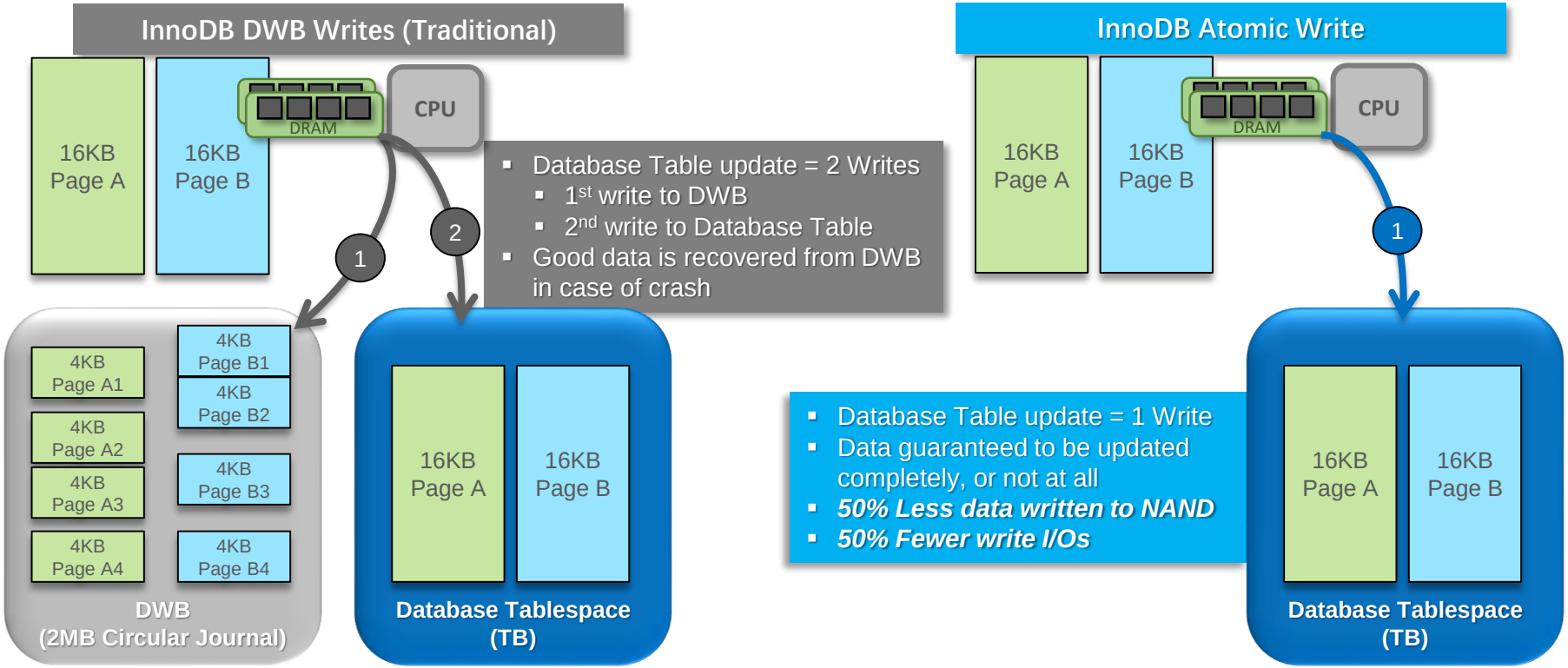
See Thomas McCormick's preso from FMS 2016 for detailed WA vs OP:
https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2016/20160809_FC12_%20McCormick.pdf

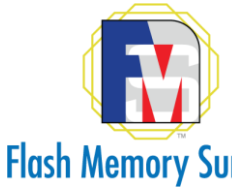
Atomic Write
Storage Driver Integration

- 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/InnoDBata unit is in one 16kB page

*SysBench Write-Only benchmarks

Atomic Write
Storage Driver Integration



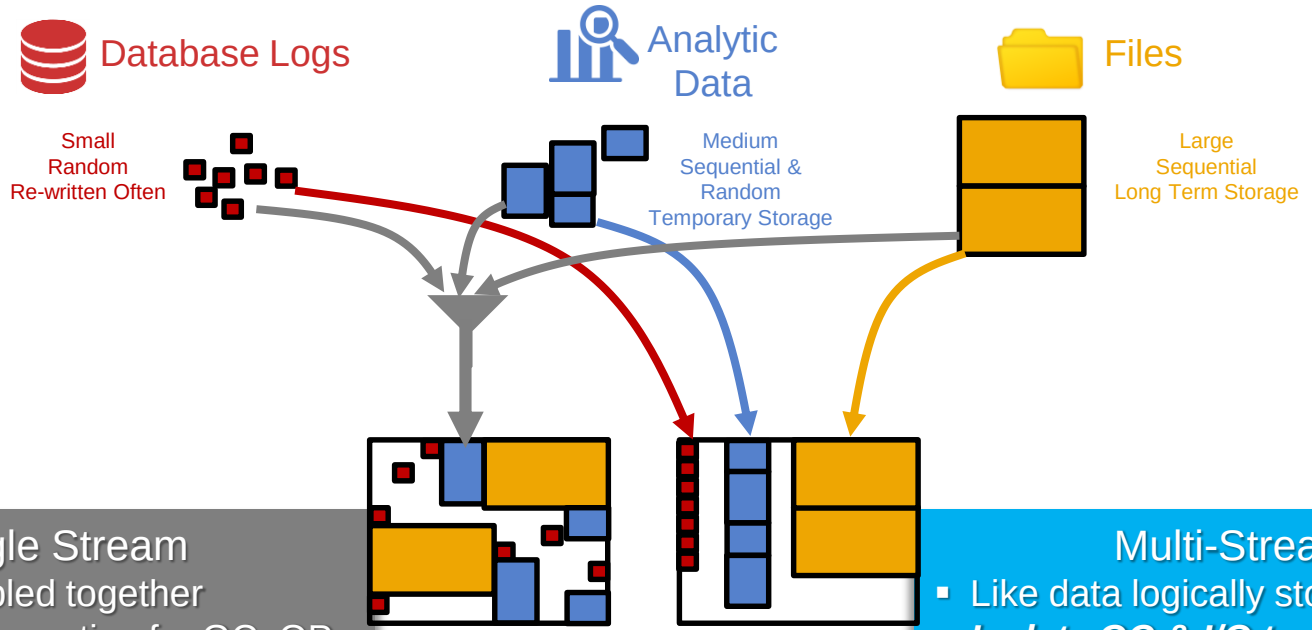


Streams
Storage Driver Integration

- 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



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

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

Flash Memory Summit

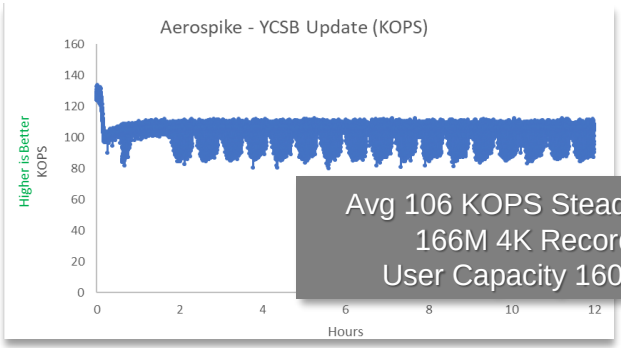
Group Garbage Collection
Application Integration

- What it is
  - Consolidation of Garbage Collection activities between the application and the SSD
- What benefits it can deliver
  - Eliminates redundant GC → Reduction in WA
  - Higher throughput & less latency variability
  - Zero-OP SSD → 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

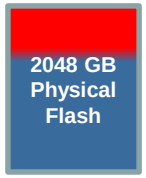
Flash Memory Summit

Group Garbage Collection
Application Integration

Baseline
Separate Defrag & GC

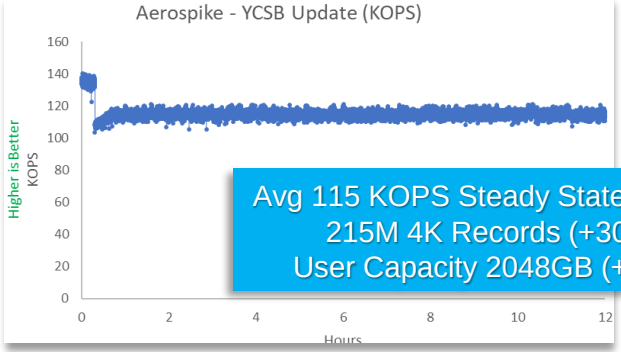


Avg 106 KOPS Steady State
166M 4K Records
User Capacity 1600GB



448GB
OP
1600GB User

0% OP Flash Storage
Group Defrag & GC



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

2048GB User

WA Reduction improves:
Endurance, Capacity, Consistency, Performance

- POC Results with CSS 1000
- Modifications to Aerospike group defragment

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