Overprovisioning in All-Flash Arrays

Cost and Benefits

Bill Radke

Director of Architecture, Skyera, Inc

Overprovision in All-Flash Arrays

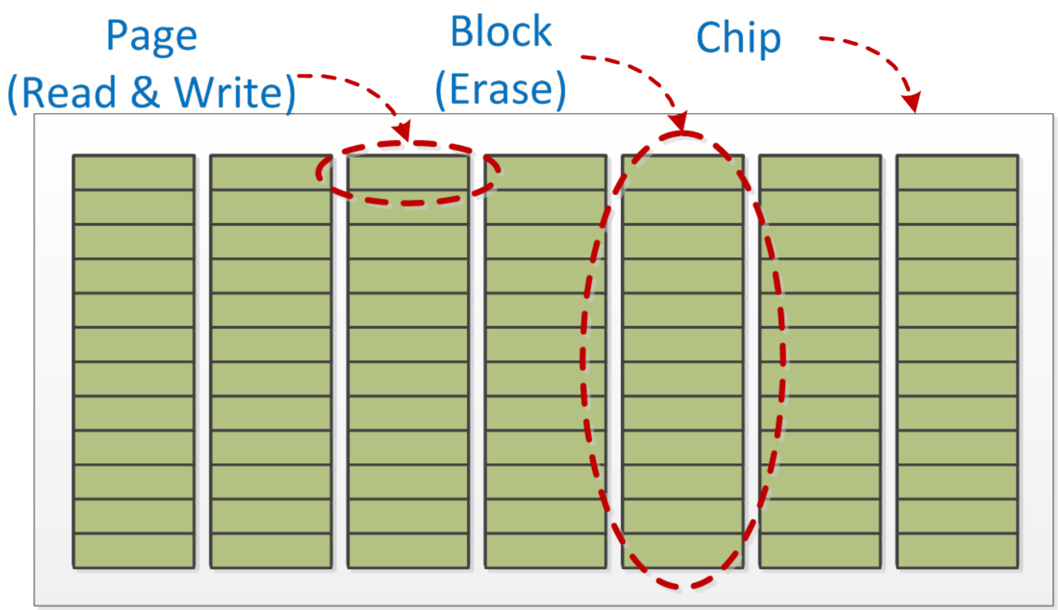
- NAND and Overprovisioning
- OP’s Effect on Performance
- All-Flash Arrays

NAND and Overprovisioning

- NAND are program/erase devices with read/write interfaces
- FTL's convert from one set of commands to another
- Overprovisioning makes that possible
  - At a cost!

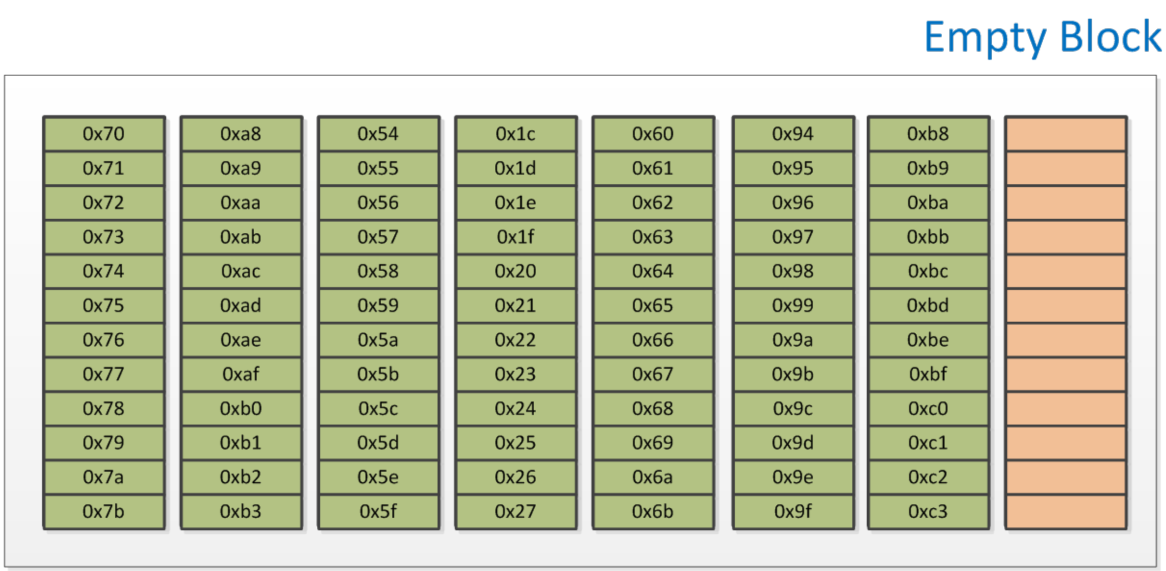
NAND & Overprovisioning: Program/Erase Devices

- Cannot overwrite data in place
- New data is written to empty blocks
- FTL notes invalidation of older entries



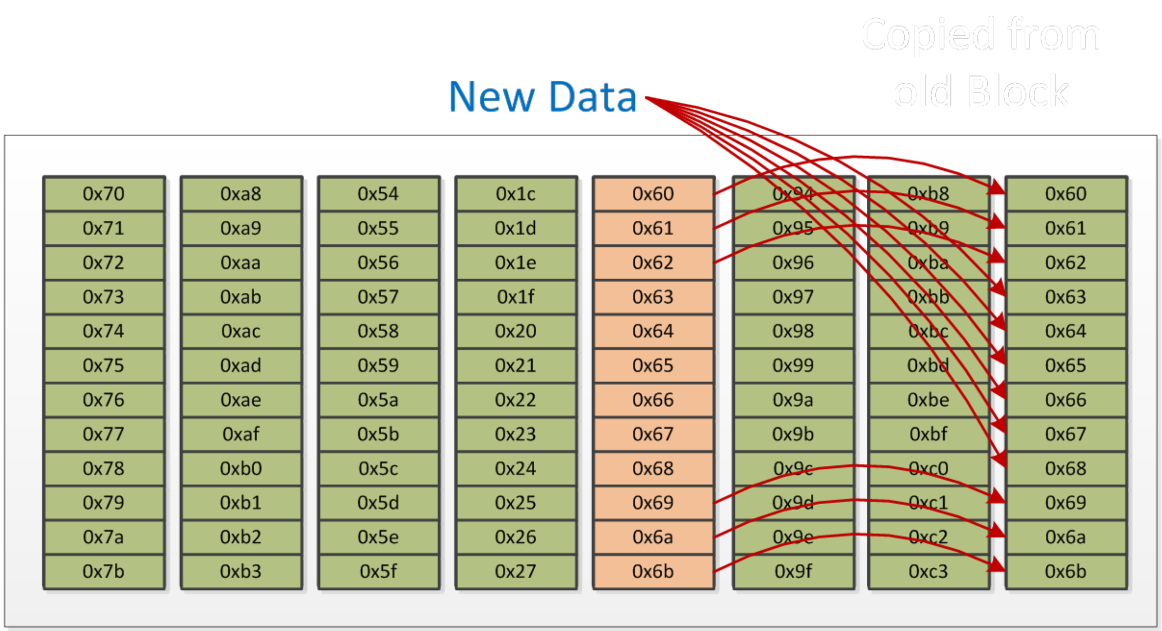
NAND & Overprovisioning: Block-based Management

- NAND was managed on a block basis
  - Any update to a given block would cause the entire block to be rewritten
- Useful if writes are at least the size of a block



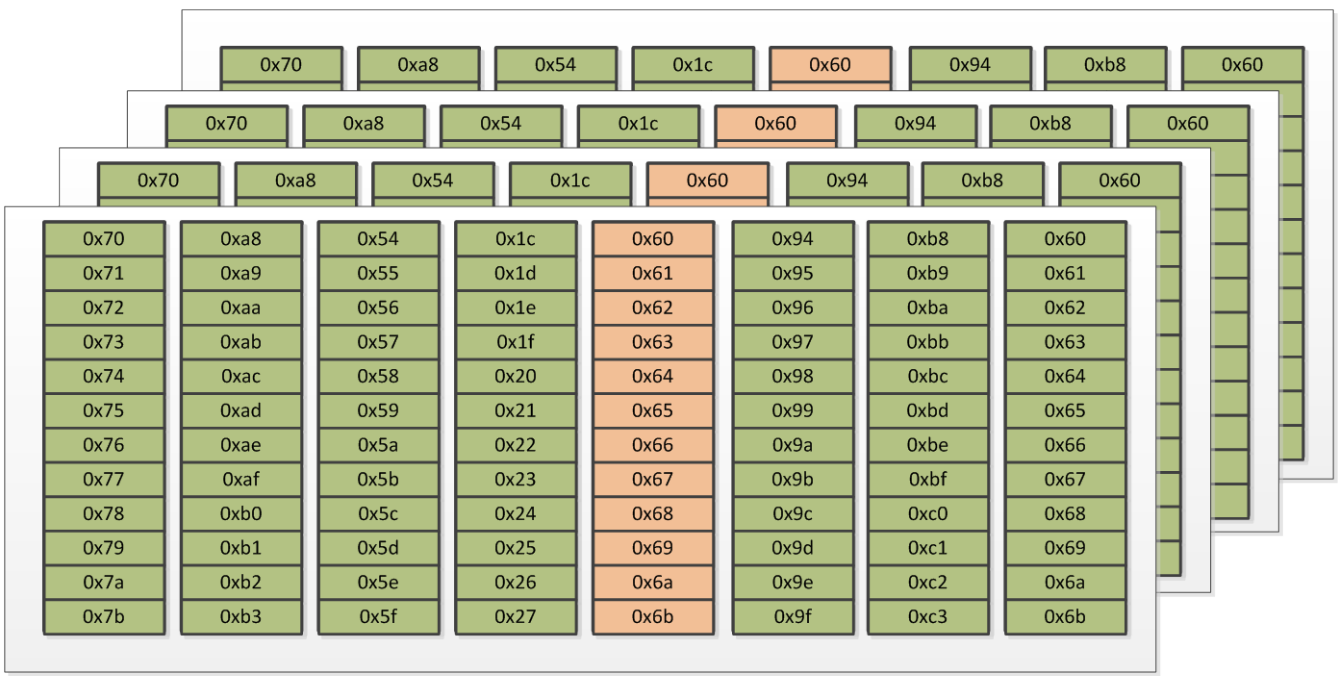
NAND & Overprovisioning: Block-based Management

- NAND was managed on a block basis
  - Any update to a given block would cause the entire block to be rewritten
- Only requires 1 spare block for OP



NAND & Overprovisioning: Block-based Management

- Becomes less efficient as the data size becomes smaller
- Becomes less efficient as blocks become bigger



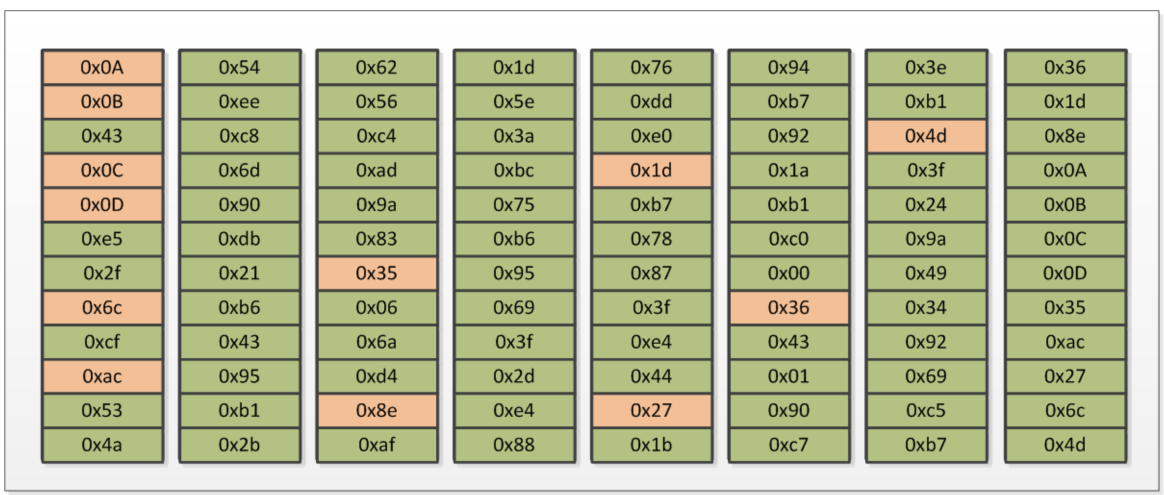
NAND & Overprovisioning: Page-based Management

- Each page of user data is tracked by the FTL
- A single block may contain many different files
- New data writes invalidate older data

0x0A	0x54	0x62	0x1d	0x76	0x94	0x3e	0x36
0x0B	0xee	0x56	0x5e	0xdd	0xb7	0xb1	0x1d
0x43	0xc8	0xc4	0x3a	0xe0	0x92	0x4d	0x8e
0x0C	0x6d	0xad	0xbc	0x1d	0x1a	0x3f	0x0A
0x0D	0x90	0x9a	0x75	0xb7	0xb1	0x24	0x0B
0xe5	0xdb	0x83	0xb6	0x78	0xc0	0x9a	0x0C
0x2f	0x21	0x35	0x95	0x87	0x00	0x49	0x0D
0x6c	0xb6	0x06	0x69	0x3f	0x36	0x34	0x35
0xcf	0x43	0x6a	0x3f	0xe4	0x43	0x92	0xac
0xac	0x95	0xd4	0x2d	0x44	0x01	0x69	
0x53	0xb1	0x8e	0xe4	0x27	0x90	0xc5	
0x4a	0x2b	0xaf	0x88	0x1b	0xc7	0xb7	

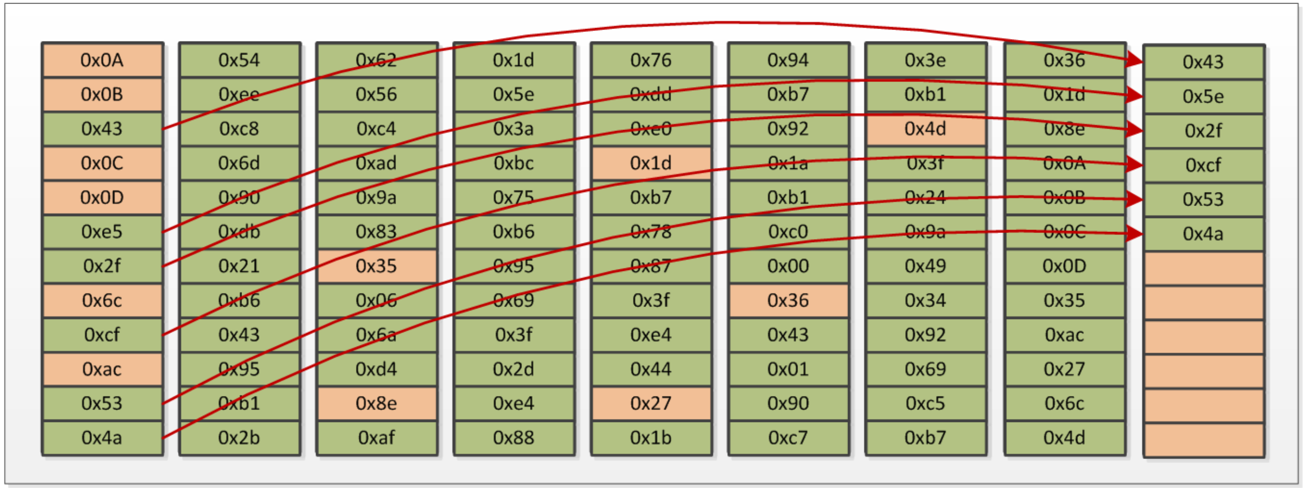
NAND & Overprovisioning: Garbage Collection

- As data is overwritten or trimmed, blocks become empty
- When a block is needed, the remaining data from the most-empty block is copied
- The block can then be erased



NAND & Overprovisioning: Garbage Collection

- As data is overwritten or trimmed, blocks become empty
- When a block is needed, the remaining data from the most-empty block is copied
- The block can then be erased



NAND & Overprovisioning: Garbage Collection

- As data is overwritten or trimmed, blocks become empty
- When a block is needed, the remaining data from the most-empty block is copied
- The block can then be erased

0x0A	0x54	0x62	0x1d	0x76	0x94	0x3e	0x36	0x43
0x0B	0xee	0x56	0x5e	0xdd	0xb7	0xb1	0x1d	0x5e
0x43	0xc8	0xc4	0x3a	0xe0	0x92	0x4d	0x8e	0x2f
0x0C	0x6d	0xad	0xbc	0x1d	0x1a	0x3f	0x0A	0xcf
0x0D	0x90	0x9a	0x75	0xb7	0xb1	0x24	0x0B	0x53
0xe5	0xdb	0x83	0xb6	0x78	0xc0	0x9a	0x0C	0x4a
0x2f	0x21	0x35	0x95	0x87	0x00	0x49	0x0D	
0x6c	0xb6	0x06	0x69	0x3f	0x36	0x34	0x35	
0xcf	0x43	0x6a	0x3f	0xe4	0x43	0x92	0xac	
0xac	0x95	0xd4	0x2d	0x44	0x01	0x69	0x27	
0x53	0xb1	0x8e	0xe4	0x27	0x90	0xc5	0x6c	
0x4a	0x2b	0xaf	0x88	0x1b	0xc7	0xb7	0x4d	

NAND & Overprovisioning: Capacity

- Extra space is provided in a number of places
  - Basically, difference between NAND and real user capacity

- Difference between base-2 and base-10
- Extra chips
- Stated capacity

NAND & Overprovisioning: Capacity

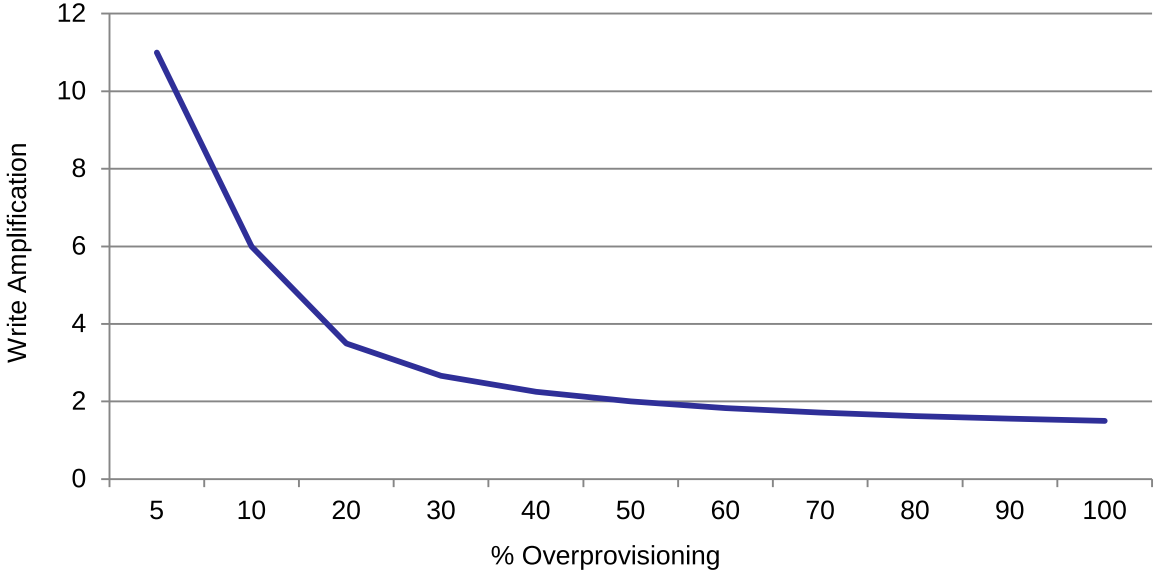
- Extra space is consumed by ECC, bad blocks, RAID, FW, tables, and OP
- Device vendors are only paid for the user capacity
  - Memory dedicated to other purposes is hidden
- Goal is to maintain performance with minimal cost

OP's Effect on Performance

- Write performance is dependent upon amount of Garbage Collection
  - With 10% OP, a system may copy 5B of GC data for every byte of write
- More OP decreases GC writes
  - Expensive way to increase write rate
- OP does not effect read performance

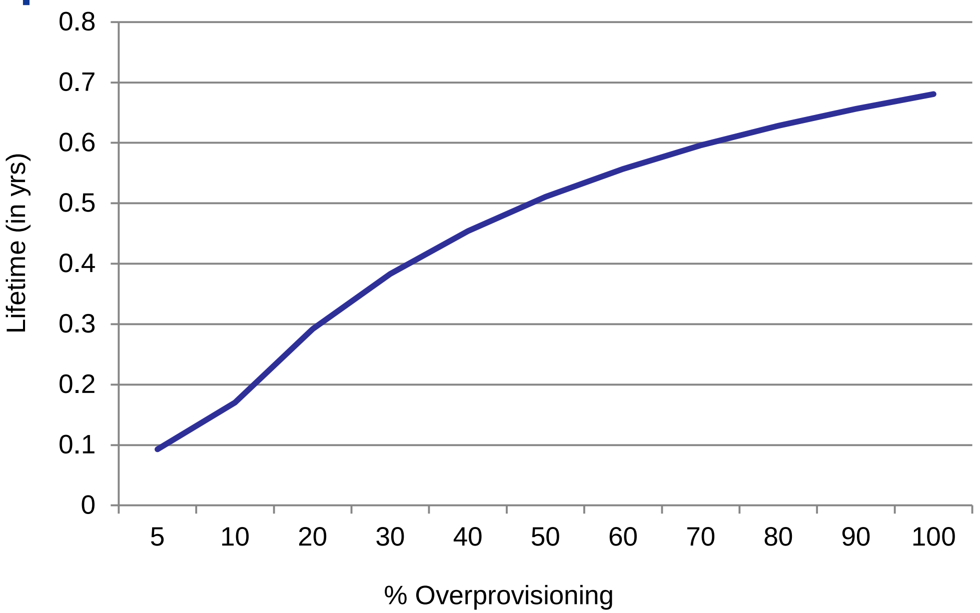
OP's Effect on Performance: Write Amplification

- Until about 50% OP, most writes are caused by Garbage Collection
- These are worst-case benchmark results
  - Small-sized, random writes



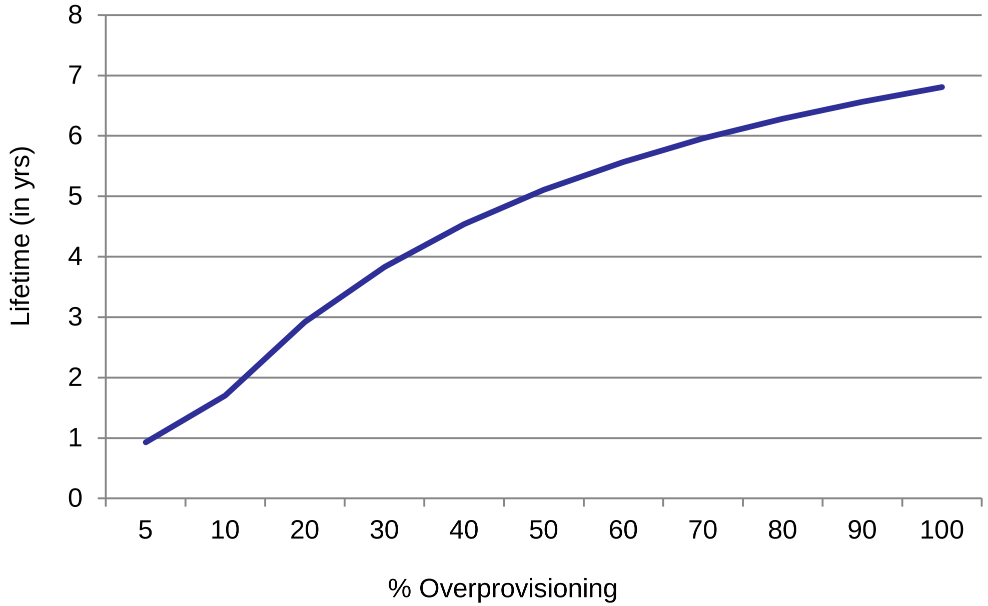
OP's Effect on Performance: Lifetime, cMLC

- A flash can be cycled to death without careful system management
- Below is 1MIOPS on a 44 TB box
- Skyera has 100x Life Amplification
  - Required to meet lifetime with cMLC



OP's Effect on Performance: Lifetime, eMLC

- A flash can be cycled to death without careful system management
- Below is 1MIOPS on a 44 TB box
  - Extensive processing not necessary for eMLC



OP’s Effect on Performance: Data Size

- Larger data sizes make page-based GC more efficient
  - Increases the chances of an entire block being invalid
- The toughest case is isolated, page-sized operations
- Random operations over a limited range are similar to large blocks

OP's Effect on Performance: Compression & Dedup

- Increases the amount of effective OP
- Radically decreases amount of GC
- Radically increases the system complexity
  - Hardware required for comp, dedup
  - Firmware to track byte locations, sizes
  - Lots of nasty edge cases for both

OP's Effect on Performance: Data Organization

- GC cost can be significantly decreased if it is not random
  - If data which is invalidated together is grouped, less copying is needed
- Many heuristics exist to improve performance
  - Based on traffic patterns of applications

OP's Effect on Performance

- A certain OP is required for a certain lifetime or performance
- The OP overhead can be based on the worst-case scenario
- The OP overhead can be informed by the transfer size, data compression and compressibility, and organization

All-Flash Arrays

- Traditionally SSD's have been used for caches in hybrid system
  - Caches are very high traffic
  - Traffic is filtered before caching
- All-Flash Arrays "see" more of the traffic
  - Different applications will have different traffic
  - There are (almost) always patterns which can be exploited

All-Flash Arrays: Traffic Characteristics

- Most applications do not require maximal OP
- Some hybrid-caching would be worst-case

Application	Xfer Size	Writes	Comp
Hybrid Caching	Small	High	Medium
Video Streaming	Large	Low	Low
Web Search	Medium	Medium	Medium
Database Processing	Small	Medium	Medium

All-Flash Arrays: Performance Tradeoffs

- The amount of OP should be kept as small as necessary
  - For both performance and endurance
- That necessary amount is smaller in All-Flash Arrays than in hybrid caches
  - Fortunate, since memory is larger

Conclusions

- Overprovisioning is a necessary response to NAND die architecture

- More is better, if everything else is equal
  - More is more expensive

- Know your application