Write Cliff Causes and Mitigation Techniques

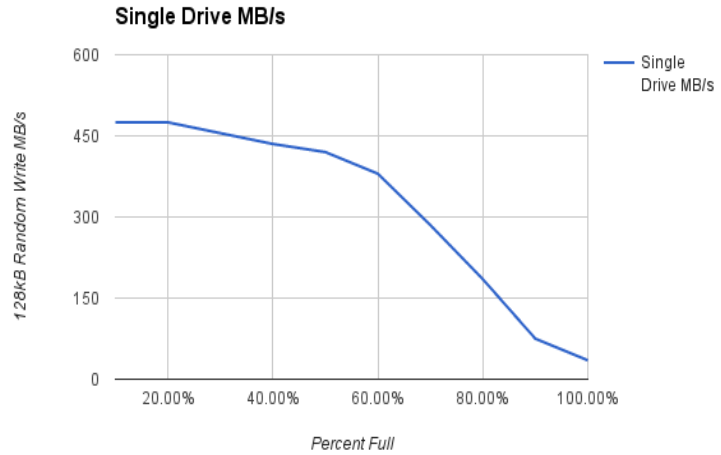
Erich Hanke
IntelliProp Inc.

Overview

- What is SSD Write Cliff?
- What causes Write Cliff?
- What types of SSD are affected?
- How can we manage the symptoms?

What Is Write Cliff

- Rapid drop in performance as an SSD is filled with data:



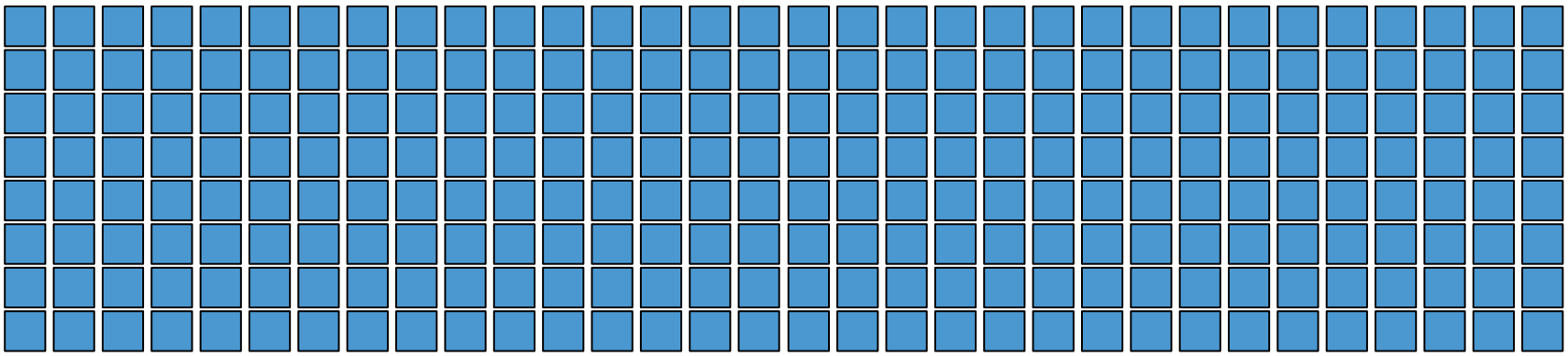
What Causes Write Cliff

Two fundamental reasons Write Cliff Occurs:

1. NAND Flash program size != Erase Size
2. NAND Flash is damaged by P/E cycle

Program size != Erase Size

- NAND Media must be erased before programmed
- Granularity of program is a “page”
- Granularity of erase is a block
  - One block is 128 to 768 pages in size

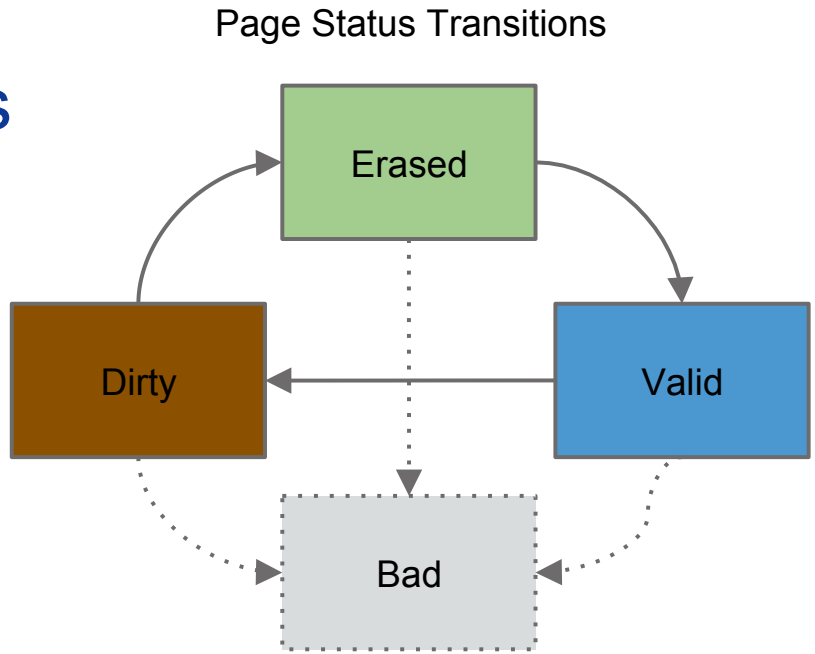


Nand Is Slowly Damaged By P/E

- The act of programming and erasing a block damages the media
- Wear should be spread evenly across the media to maximize drive life
- Wear Leveling algorithms must be employed by the SSD controller to spread wear
- Garbage Collection algorithms must be employed by the SSD controller to track “in-use” vs. “not-in-use” pages and blocks

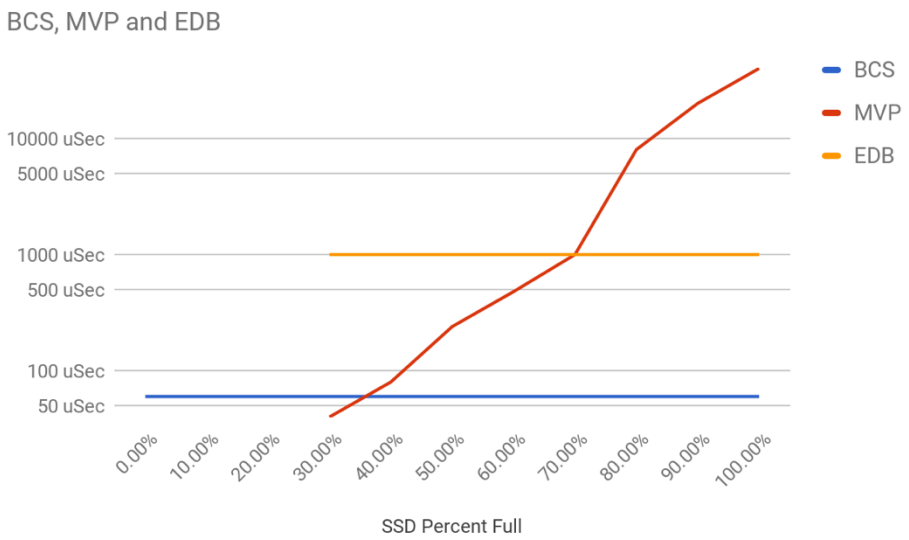
Garbage Collection

- Goal: Make “Clean” Blocks
  - Relocate “Valid” pages
  - Erase “Dirty” blocks
  - Make “Clean” blocks



Garbage Collection Tasks

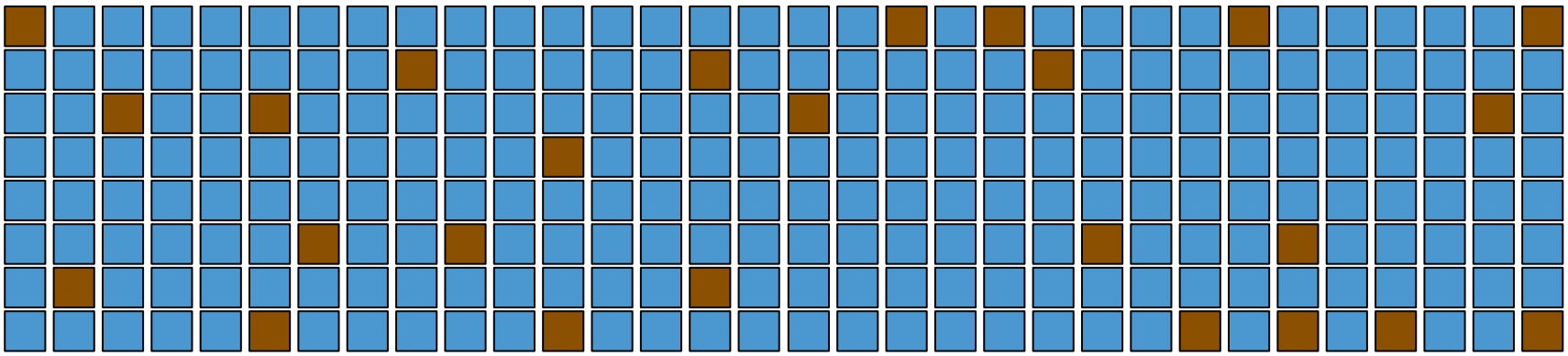
- Search for block candidates
  - Move “valid” pages
  - Erase “dirty” blocks



* Assuming 7% OP

Finite NAND Bandwidth

- These Garbage Collection are NAND specific, not interface specific
  - Same NAND tasks for USB, SATA, SAS, NVMe, etc.



Mitigation Techniques

- Increase Over Provisioning (OP) percentage ($$$)
- Increase Controller DRAM Buffer/Cache ($$)
- Use RAID to distribute host load ($ - $$)

Increasing Over Provisioning

- The Good:
  - Decreases average work done on a given page by GC engine
  - This technique certainly improves Write Cliff symptoms
- The Bad:
  - Increases the $/GB cost of the SSD
  - Potentially limits the max SSD capacity

Increase Controller DRAM

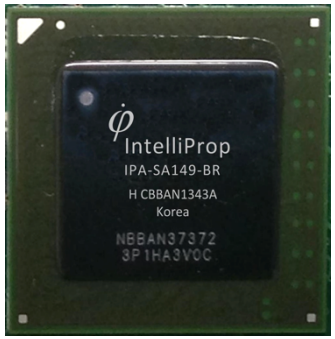
- The Good:
  - Allows host activity to be buffered while GC activity is ongoing
- The Bad:
  - Complicates unexpected loss of power handling
  - Increases SSD Power
  - Merely postpones the performance droop during heavy write workloads

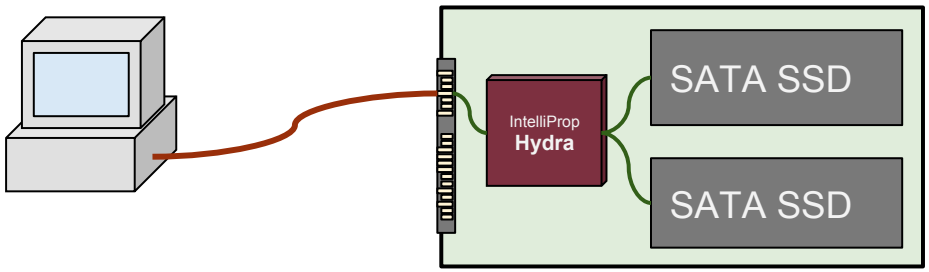
RAID Host Load Distribution

- The Good:
  - Host bandwidth distribution across multiple SSD can minimize application observed performance droop
- The Bad:
  - RAID-5 / 6 makes WC effects worse by increasing write activity
  - Adding parity increases cost per usable byte
  - Many RAID controllers are unable to issue “TRIM” commands
    - No TRIM support “Locks in” the performance droop

IntelliProp "Hydra" Bridge

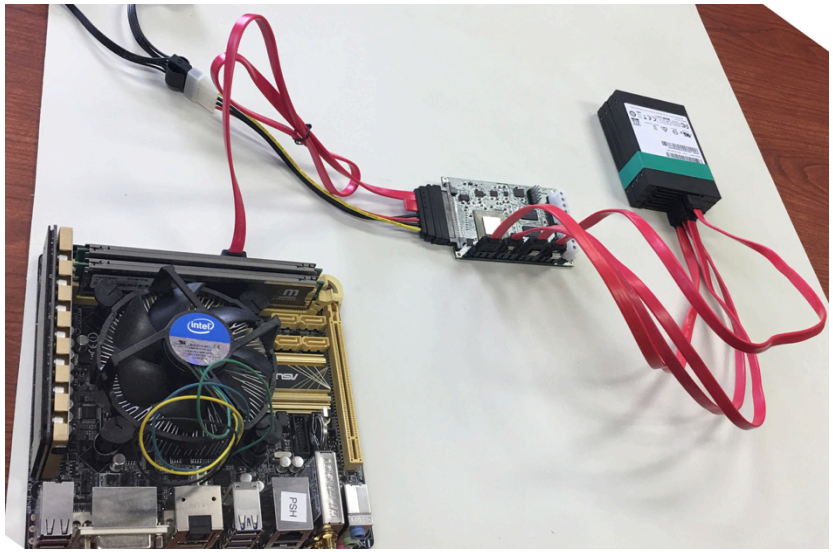
- Supports Data Set Management (TRIM)
- Hardware stripping engine
- Automatic Host Load Distributing
- Multiple Configurations
  - 1:2, 1:4, Cascade support



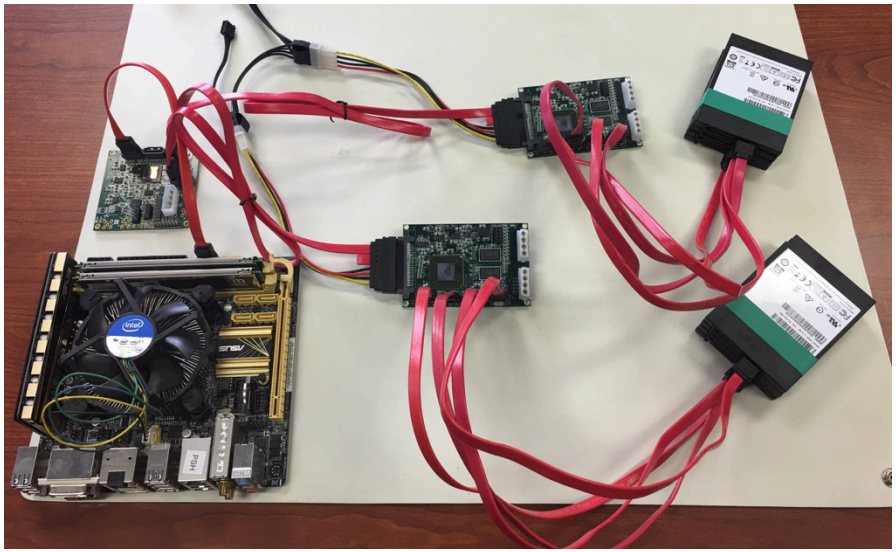


Write Cliff Testing

Hydra 1:4



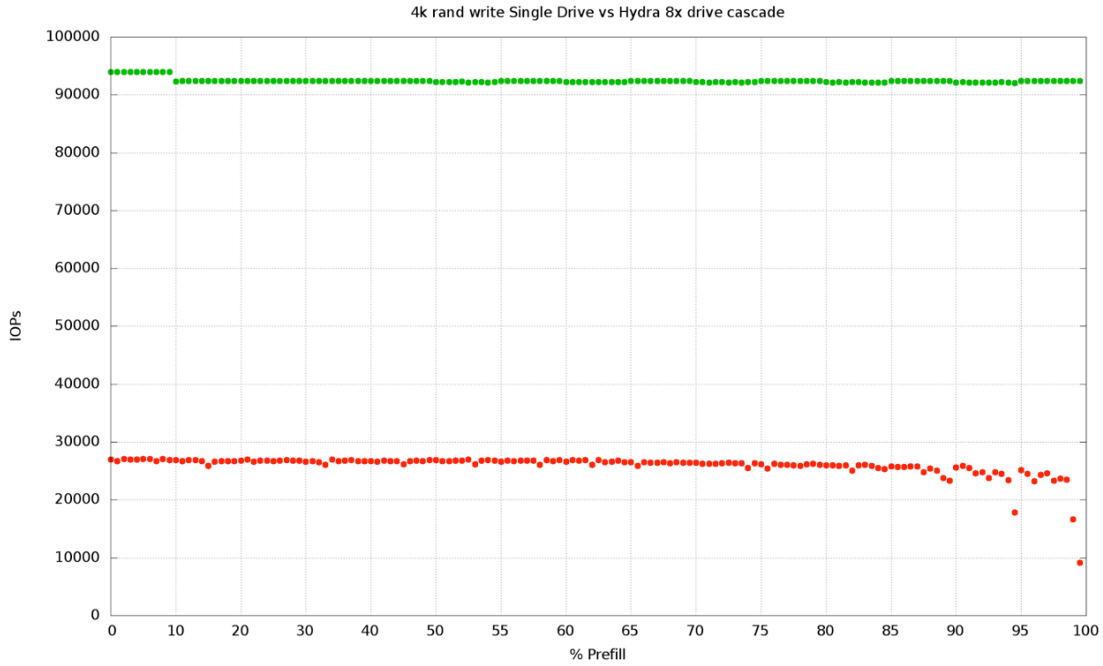
Hydra 1:8
Cascade



Test Procedure

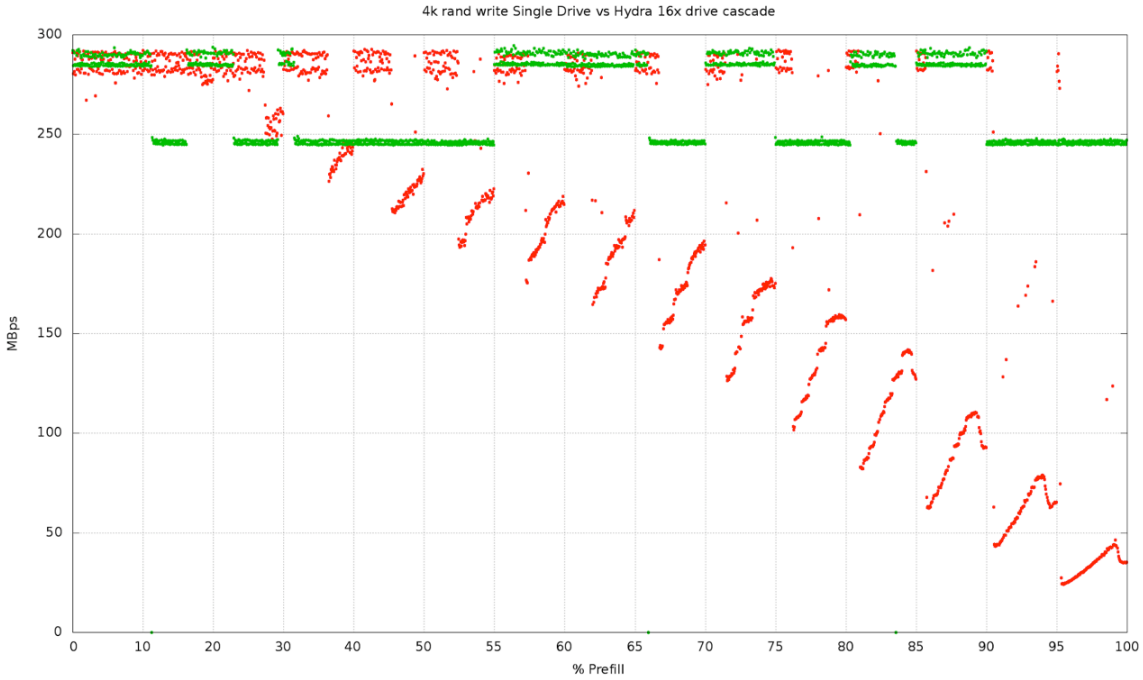
1. Secure Erase Drive(s)
2. Run Automated pre-conditioning script to fill SSD to a particular LBA% used
3. Run stepped iometer testing script
4. Increment LBA% and go-to step #2
5. Once the script completes the LBA%-100 step, testing complete

Results (Enterprise MLC SSD)



- Above 80% full, the enterprise SSD starts to suffer IOPS loss. Eventually dropping to 50% of peak performance.
- Hydra (using same SSD) stays within 5% of original performance across all fill points.

Results (TLC SSD)



- Single drive bandwidth drops by over 90% from erased to full!
- Hydra (using same SSD) stays within 20% of original performance across all fill points.

Summary

- SSD slow as the percentage of used LBA increases to 100%
- These effects occur regardless of interface type (SATA, SAS, NVMe)
- These effects can be managed

Thank You

- References
  - "IntelliProp Hydra Technology Mitigates SSD Write Cliff"
  - www.intelliprop.com/support-white_papers.htm
- More Information
  - Intelliprop Inc.
  - ehanke@intelliprop.com
  - www.intelliprop.com