Storage. Networking. Accelerated.™

What Can/Can't be Done and What's Nice to Have in the New Stack

Steve Peters – LSI Corporation

Contents

Overview

- SSDs share the I/O interface technology developed for hard disk drives, thus permitting simple replacement for most applications.
  - SNIA: Solid State Storage 101: An introduction to Solid State Storage

- SATA or SCSI makes it easy to adopt the technology but may not be the optimum interface.

WHAT IT TAKES TO MAKE A SSD

What it takes to make a SSD

- SSD Controllers / IP define the personality of the SSD. Controller architecture and effective implementation processes transform unreliable me-too memory chips into the diverse range of application optimized SSDs that you can see in the market today.
  - Zsolt Kerekes / StorageSearch.com

SSD Basics

- GB versus GB*
  - 2^{30} versus 10^9 is 1.0737
  - 7.37% of the Flash is Available for Metadata, bad blocks, etc. when a drive's size is the same as the available Flash
- Map Table
- Garbage Collection (aka Recycling)
- Wear Leveling
- Write Amplification
- Over provisioning
- Trim Command
- RAISE

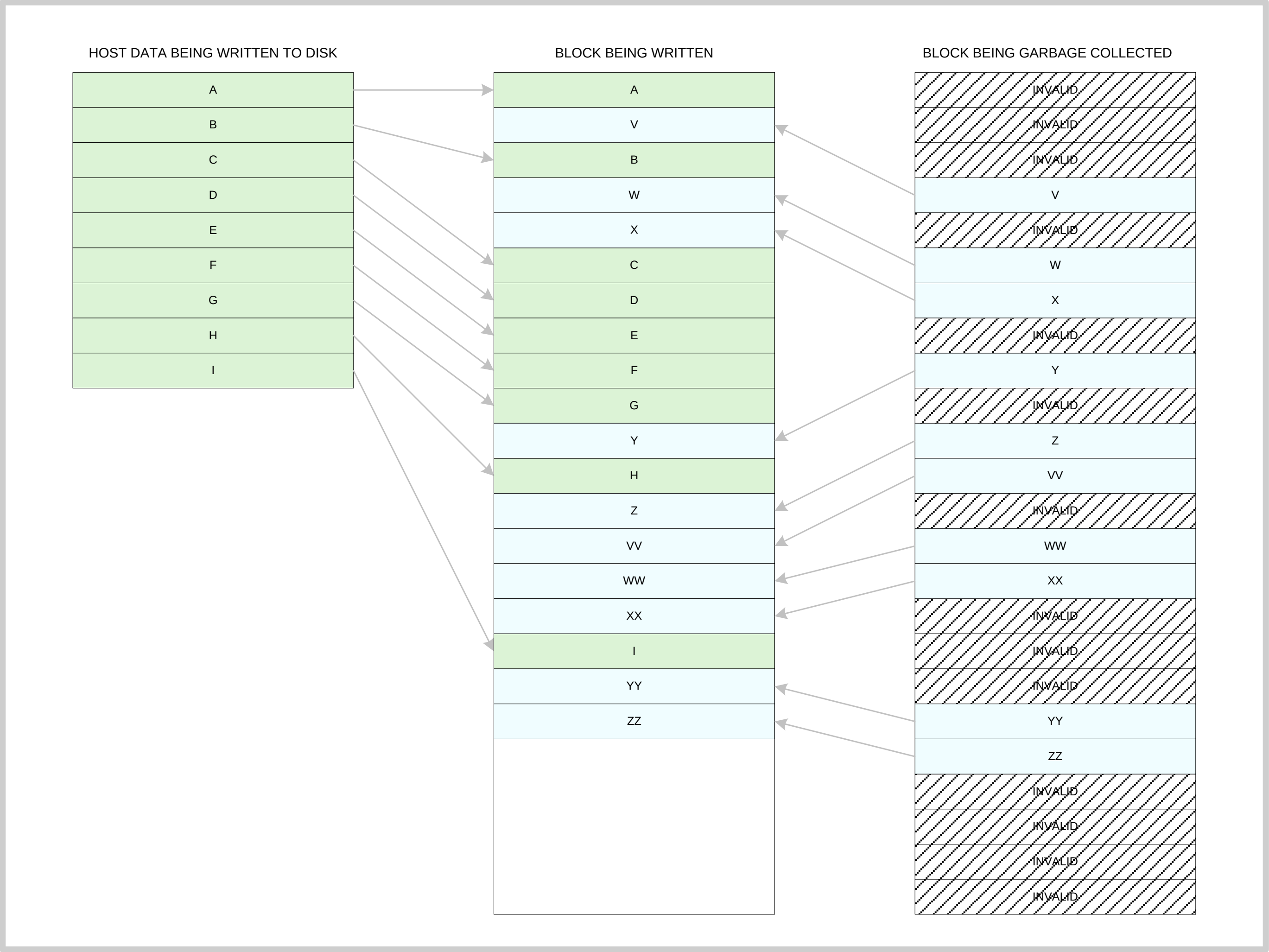
Map Table

- Sector data is accumulated into pages which are written to Flash
  - For example, 16 sectors could be written to an 8-KB Flash Page
  - Since the sectors are aggregated as they are received, a sector can be located anywhere in Flash.
  - Therefore, a mapping table needs to be maintained which contains the address of every sector.
  - The Table needs to be accessed when a sector is read to determine which Flash data needs to be read.

Garbage Collection

- Data in a block becomes free as sectors within a block are written.
  - For example, block A contains sectors X, Y, and Z. When the host writes X, the new location for X is in Block B. The version of X in block A is now “stale” and represents free data (i.e. garbage).
- Garbage Collection
  - When the number of “Free” Flash Blocks reaches a low level, blocks need to be freed.
    - The block with the most free data is selected and any valid data in the block is rewritten.
    - Garbage Collection can occur concurrently with data being written by the host.
- Synonymous with “Recycling”

Garbage Collection (2)



Wear Leveling

- Wear Leveling is the process of evenly distributing the p/e cycles over all of the blocks.
  - When garbage collecting, the blocks which have little or no free data will not be selected. The effect is that a set of blocks may have a low p/e count while another set may have a high p/e count.
  - The wear leveling algorithm needs to select less attractive blocks when the delta p/e count (highest p/e count – lowest p/e count) crosses a threshold.
    - This may include blocks that have no free data.
    - Should interleave “good” blocks (high free count) and “bad” blocks (low free count).

Write Amplification

- Write Amplification
  - The Ratio of data written to the Flash Memory divided the amount of host data written to the drive.
    - Ideal value is 1.0
  - Write amplification occurs due to the need to perform garbage collection and wear leveling.
    - The data rewritten during garbage collection is in addition to the host data (i.e. the write data is more than 1x what the host wrote).
    - The amount of data rewritten by the garbage collection process will vary depending on several parameters:
      - Type of Host Accesses (Sequential or Random)
      - Amount of Overprovisioning
      - State of the drive
      - Etc.
  - Reduces host write performance
    - Flash bandwidth used for writing of recycled data.
    - Reduces drive life
  - Results in more p/e cycles

Overprovisioning

- Overprovisioning
  - Providing more Flash Memory than the specified drive size.
    - For example, using 128 GB of Flash for a 100 GB* drive results in 28% overprovisioning.
    - A drive with the same amount of Flash as the drive size is considered to be 0% overprovisioned.
    - Reduces write amplification
      - Blocks will have more free data.

Trim Command

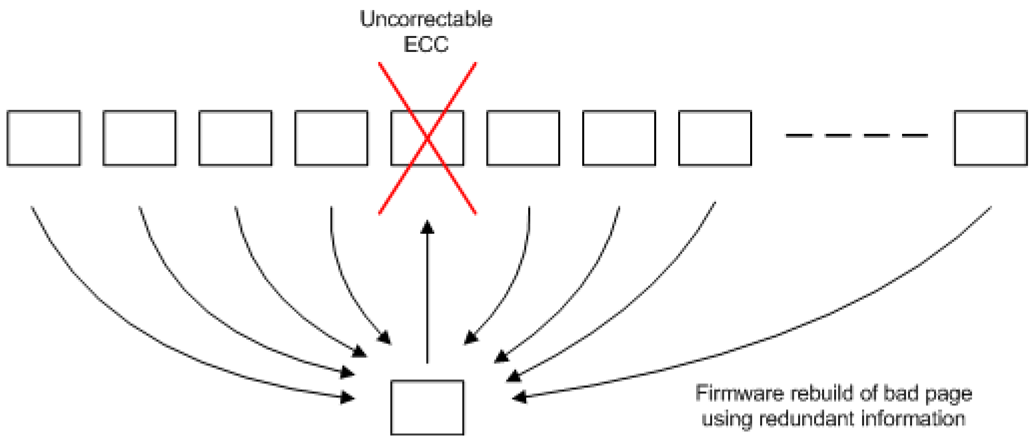
- Trim
  - Allows Host to provide guidance to an SSD on areas of the drive that the Operating System no longer considers valid.
    - e.g. Files removed from the wastebucket
    - Host provides starting LBA and length for trimmed region.
    - SSD can mark the sectors in the trimmed region as “free”.
    - Provides a similar effect to overprovisioning
      - e.g. a 256 GB laptop drive has only 100 GB used
        - If the other 156 GB is trimmed, the SSD is effectively a 100 GB drive with 156 GB of overprovisioning (at least temporarily)

Compression

- LSI SandForce FSPs compresses the host data to reduce the amount of data written to Flash.
  - Less Data written to Flash
  - More data in blocks will be free reducing recycled data

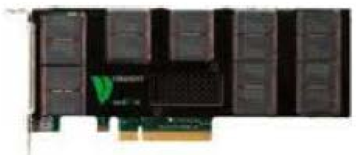
- Compression is over 4KB blocks
  - Compression over 512B is less efficient
    - Therefore, internal sector size of 4K is used.

RAISE



- The unique LSI RAISE™ (Redundant Array of Independent Silicon Elements) technology. RAISE technology provides the protection and reliability of RAID on a single drive without the 2x write overhead of parity.
- On Uncorrectable ECC error, Firmware uses redundant information in multiple pages to rebuild bad data.
- RAISE recovery takes about 30ms
- Recovered data is rewritten.

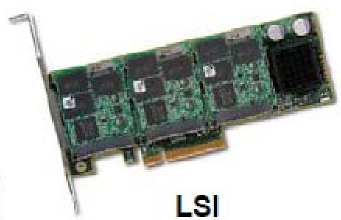
Enterprise SSDs



Virident

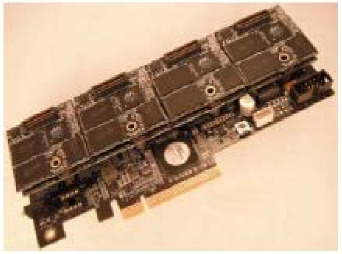
Fusion-io

Micron



LSI

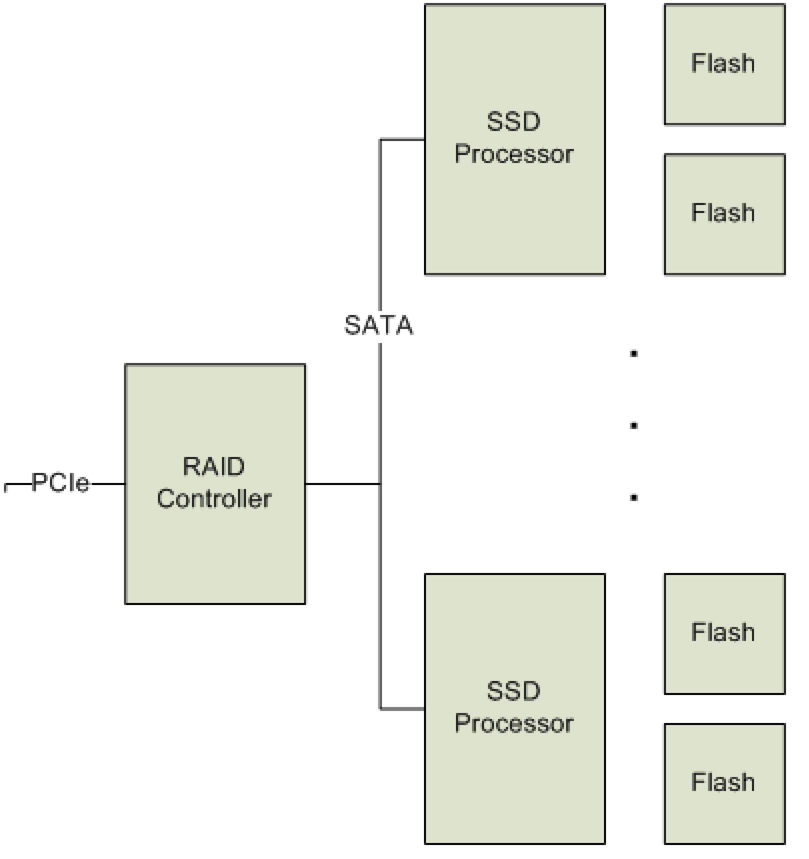
OCZ



Marvell

Enterprise SSD

- PCIe frontend
- RAID chip
  - process commands
  - Distribute I/O
- Multiple SSD processors
  - Implements FLASH functions
- Uses on-board SATA as device interconnect
  - Allows for pluggable daughter boards



NEAR FUTURE ENHANCEMENTS

Coming soon or Already here

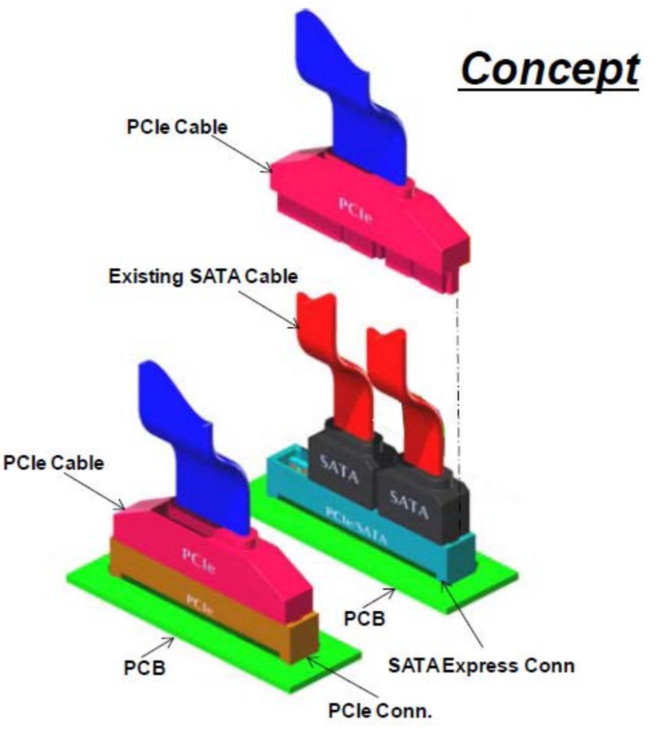
- Next Generation SSD reduces latency and increases bandwidth
  - Continues to use block I/O model

- PCIe interconnect for SSD processor

- NVMe protocol

PCIe

- PCIe Interface for SSD processor
  - Provides higher bandwidth interface
  - Enterprise SSDs can use PCIe to interconnect multiple SSD processors
  - Limited by PCIe slots available on systems
- SATA Express
  - PCIe interconnect for devices



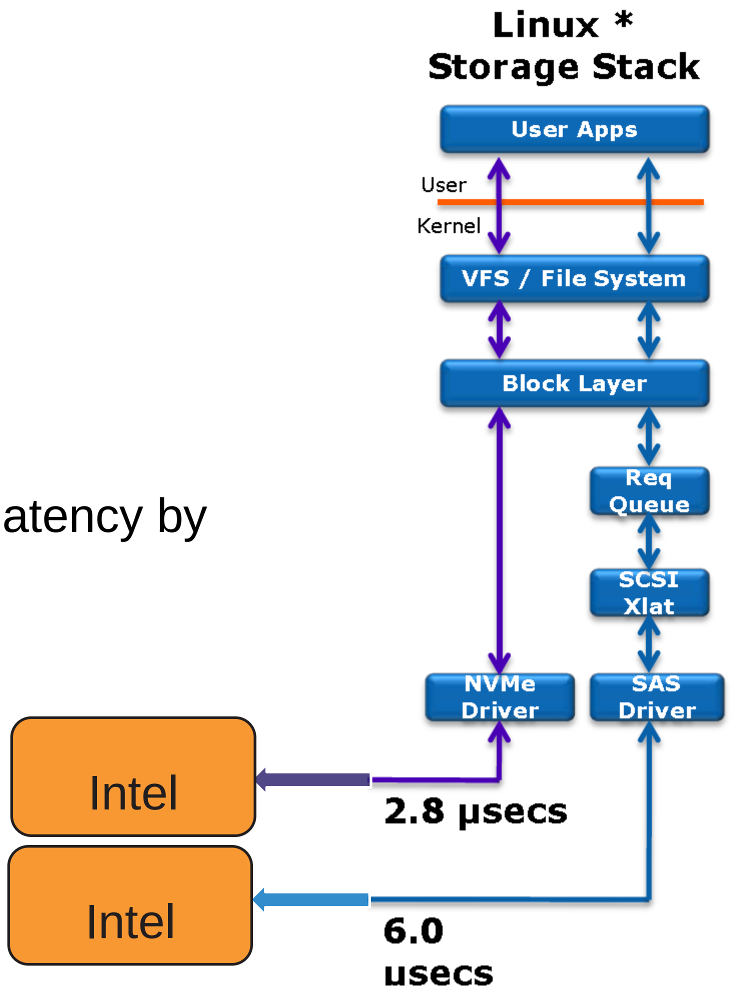
NVME Driver stack

- NVMe protocol
  - Designed for direct PCIe interconnect
  - Smaller more efficient commands
  - Lower latency
  - Supports multiple command queues

  - According to Intel's study NVMe reduces the latency by more than 50%.

    SCSI/SAS: 6.0 usecs 19500 cycles

    NVMe: 2.8 usecs 9100 cycles



WHAT NVM PROGRAMMING IS ASKING

NMV Programming

- Defines a new model for accessing NVM.
  - The model for Persistent Memory shares most behavior with the native OS's model for memory-mapped I/O.

  - Operations
    - Allocate Region – reserve NVM space for use by NVM objects
    - Create objects – create the object and reserve NVM space
    - Open – gets handle to NVM object
    - Map - maps NVM space to the application's address space.
    - Un-map – synchronizes , then releases the mapping.
    - Sync – forces a synchronization between memory and NVM space.
    - Close – Synchronizes, Un-maps, and releases the object handle.

NMV Programming – Hardware requirements

Features

- Regions – may use the same firmware feature as the NVMe namespace function.

- Objects – will require meta data space to be reserved and a simple object store data structure to be implemented. This will also interact with the SSD translation layer.

- Map - maps NVM space to the application's address space.
  - Must have some way to track what has changed.
  - DMA will need to read scattered blocks.
  - The SSD translation layer will need to interact with non contiguous lists of blocks to write.

Does NVM programming model help?

- Cons:
  - To make NAND flash useable the SSD processor will still need to perform all the functions it does today.
  - To use the new model software will need to be re-written.

- Pros
  - By mapping to object to memory, a large number of scattered blocks can be made to update efficiently using a single sync command.
  - With a little extra work the sync can be made atomic.

LSI®

Storage. Networking. Accelerated.™

SandForce DRIVEN™ By LSI

Follow SandForce