CRITICAL I/O WHITEPAPER

SSD Flash Technology Options

for StoreEngine and StorePak

Abstract

StoreEngine and StorePak are designed to leverage standard form factor SSDs. There are many SSD options including SLC, eMLC, pMLC, MLC, and TLC, and much resulting confusion. The application note compares the attributes and resulting reliability of the base flash technology options in the context of StoreEngine and StorePak.

Revision 2/9/2015

Flash Technology Choices for StoreEngine and StorePak

This paper will discuss issues related to flash technology and SSD selection for Critical I/O's StoreEngine and StorePak products. The paper is organized as follows:

1. Introduction.
2. MLC vs. eMLC vs. SLC flash technology overview.
3. Comparing the SSD flash technology alternatives.
4. SSD and flash controller architectures for MLC Flash.
5. What type of flash is the best fit for my application?

1. Introduction

StoreEngine and StorePak are designed to leverage industry standard form factor SSDs. There are many SSD options, and much resulting confusion. Standard SSDs are generally based on one of three main types of flash:

- **MLC** – Stores two or more data bits per cell, highest density and performance
- **eMLC** – MLC flash enhanced for better endurance and retention
- **SLC** – Stores one data bit per cell

Because MLC and eMLC flash store two or more bits of data per flash cell, they have significant cost, density, and power advantages over SLC flash. These advantages do come at the cost of reduced write endurance and/or reduced data retention. Whether this reduction is significant is determined by the requirements of the intended application. The characteristics of many systems are such that MLC flash is perfectly well suited, but certain system applications may demand the use of SLC flash.

The unavoidable characteristics of all flash media are: 1) Limited write endurance (wear-out), and 2) Limited power-off data retention. While different flavors of flash (such as SLC and MLC) vary in the specific numbers attached to these characteristics, all varieties of flash are affected by these two characteristics.

Note that it is important to differentiate between SSD "*failure*" and "*wear out*". SSDs are highly reliable, much more so than hard drives, and a full SSD *failure* (akin to a hard drive crash) is an exceedingly rare event. Typical SSD MTBFs range from 1.2M to 2.0M hours. SSD *wear out* (due to the SSDs reaching their block erase/program count rating) is a highly predictable phenomenon that can be monitored and does not sneak up on a user.

As a full SSD failure is quite rare, RAID techniques that are commonly applied to hard drive based systems are not as applicable to SSDs. In particular, the use of RAID does not offer any improvement to SSD endurance related wear-out; in fact, the increase in SSD writes associated with RAID (due to added mirror or parity writes) can actually reduce the lifetime of an SSD based system.

Critical I/O's StoreEngine and StorePak products provide a SSD wear out indicator which can be monitored to provide the "wear out" status of the individual SSDs. In environments where this is some risk of SSD wear out, this provides a method to proactively refresh SSD storage before any performance degradation occurs.

Critical I/O SSD Nomenclature

The specific SSD types used in Critical I/O storage products are designated by a suffix that is appended to the basic StoreEngine or StorePak model number.

The suffix is of the form –XYzzzz, where:

- X indicates the basic flash technology (S = SLC, E = eMLC, M = MLC, T = TLC)
- Y indicates the application focus (C = recording applications, S = server NAS/DAS applications)
- zzzz indicates the raw storage capacity in GB (the usable capacity is always slightly less)

Examples:

- SP306R-MC3072 - MLC, oriented towards recording applications, 3072 GB raw capacity
- SP306R-ES2880 - eMLC, oriented towards server NAS/DAS applications, 2880 GB raw capacity

SSDs that are optimized for recording applications provide higher and more consistent large block sequential write performance (i.e. focus is higher MB/s). SSDs that are optimized for server NAS/NAS applications provide better and more consistent small block random read/write performance (i.e. focus is higher IOPS).

2. MLC vs. eMLC vs. SLC Flash Technology Overview

The table below compares the attributes of the base flash technology options in the context of StoreEngine and StorePak. The data presented is typical for the SSDs used in StoreEngine and StorePak configurations, but the precise data values will vary according to the specific SSD model and manufacturer.

Table 1. SSD Flash Technology Options (typical, for a single SSD)

Single SSD	**MLC**	**eMLC**	**SLC**
Capacity (current)	1000 GB	800 GB	400 GB
Capacity (est Q4 2015)	2000 GB	1600 GB	500 GB
Relative Density (est Q4 2015)	1x	0.8x	0.25x
Relative Cost	1x	1.5x	8x
Relative Power (per byte)	1x	1.1x	5x
Operating Temperature Range	-40 +71	-40 +85	-40 +85
Typical Write Endurance	4,000	20,000	60,000
Typical Write Performance	500 MB/s	400 MB/s	250 MB/s
Single Drive MTBF	1,200,000	1,500,000	2,000,000

These flash technology options are briefly described below in order of increasing cost per bit. Note that "cost" is not only measured in price; rather it is also measured in watts, weight, and size. Using the highest density flash technology that is compatible with the application requirements will generally result in the lowest cost solution, as measured in terms of price, power, weight, and size.

MLC Flash (2 bits/cell) – MLC flash offers the highest capacity, the highest sequential write performance, and the lowest cost per GB. MLC drives have ~3x the capacity at 1/8 the cost (on a per GB basis) of SLC for equivalent packaging methods. This disparity will increase, as nearly all industry flash research dollars are going into MLC flash development. A special case of MLC flash is TLC flash with 3 bits/cell. *Due to limited endurance and limited temperature range, TLC flash media is suitable only for selected applications.*

Critical I/O products that use MLC flash are designated by a –MC or –MS suffix (example: -MS3072). Critical I/O products that use TLC flash are designated by a –TC suffix (example: -TC6000).

eMLC Flash (2 bit/cell) – Enhanced MLC or eMLC flash results from a tuning of standard MLC processes along with more sophisticated controller and error management technologies that provides improved write endurance, data retention, and thermal performance as compared to standard MLC. This comes at the expense of increased cost and often slightly reduced capacity (due to a higher level of overprovisioning).

Critical I/O products that use eMLC flash are designated by a –EC or –ES suffix (example: -ES3072).

pMLC Flash (1 bit/cell) – Pseudo MLC or pMLC flash is the result of using MLC flash chips, but using them in a 1 bit per cell mode. This provides endurance and retention characteristics are similar to those of eMLC flash but at only one half of the density of MLC.

SLC Flash (1 bit/cell) – Single Level Cell flash offers the best write endurance, and slightly better data retention and thermal performance. This comes at the expense of much higher cost and reduced capacity. Recommended for applications where only the highest data reliability is required and the associated reduction in storage capacity and increased cost is acceptable

Critical I/O products that use SLC flash are designated by a –SP suffix (example: -SP1200).

Table 2 shows the performance and capacity attributes for StoreEngine and StorePak using SSDs based on the three main types of flash technology.

Table 2. StoreEngine and StorePak Characteristics vs. Flash Technology

6U VPX StoreEngine (3 SSDs)	**MLC**	**eMLC**	**SLC**
StoreEngine Capacity (current)	3.0 TB	2.4 TB	1.2 TB
StoreEngine Capacity (est Q4 2015)	6.0 TB	4.8 TB	1.5 TB
Typical StoreEngine Write Performance	750 MB/s	750 MB/s	750 MB/s

6U VPX StorePak (6 SSDs)	**MLC**	**eMLC**	**SLC**
StorePak Capacity (current)	6.0 TB	4.8 TB	2.4TB
StorePak Capacity (est Q4 2015)	12.0 TB	9.6 TB	3.0 TB
Typical StorePak Write Performance	2500 MB/s	2000 MB/s	1800 MB/s

3. Comparing the SSD Flash Technology Alternatives

The factors to consider when evaluating MLC vs. eMLC vs. SLC flash options are:

1. Performance
2. Cost and Density
3. Endurance (amount of data written, sequential vs. random writes)
4. Data Retention (related to endurance)
5. Temperature (affects power-off retention)

Performance

At the flash chip level, SLC flash offers a performance advantage vs. MLC flash. At the SSD level, however, current generation MLC based SSDs tend to offer higher sustained sequential read and write performance than SLC based SSDs. This is due to the large flash block sizes and sophisticated flash write management that are built into most MLC based SSDs.

Cost and Density

When choosing between the main categories of flash (MLC, eMLC/pMLC, and SLC), it can be tempting to immediately choose SLC, as this often appears to be the safe choice. And if maximum write endurance and retention are critical to the application, SLC may be the right choice. But one must consider that use of SLC flash comes at the cost of typical 4x reduction in density (and an associated increase in SWaP) and a typical 8x increase in price) as compared to MLC flash. (density/price as measured at the SSD level)

Because of the huge consumer demand for MLC NAND, MLC enjoys high density packaging, large investments in controller design, and manufacturing economies of scale that give it a cost per bit cost of well under that of SLC, and it is expected that this disparity will increase.

Write Endurance and Wear-Out

Flash cells “wear out” as they are erased and re-written. The number of erase/program cycles that can be performed before wear out vary with the flash type, ranging from less than 1000 cycles per flash block for TLC, from 3000 to 5000 cycles for MLC, 15,000 to 30,000 cycles for eMLC/pSLC, to over 60,000 cycles for SLC flash.

This significance of this wear-out factor depends largely on the application. For example, for recording applications, where SSDs may be filled once per mission using large block writes, the wear-out factor is normally not an issue. But for a high bandwidth storage server that performs a high volume of small writes to SSDs, the wear-out factor may become significant due to an effect termed “write amplification”. This affect most often comes about when SSDs are subjected to sustained very small random writes. Because flash is written a page at a time, and can only be erased a block (which is many pages) at a time, a sustained sequence of small random writes can result in a much higher volume of actual flash writes due to the need for data movement (internal copies) for page/block consolidation. In such usage cases, eMLC or SLC flash media may be the best media choice.

Critical I/O's storage product provide a wear out indicator that tracks SSD usage, and that can be monitored in the user's system. This provides a capability to both tracks the SSD wear out rate, as well as provide an advance indication that the SSD may be nearing the end of its useful life and should be replaced.

Endurance vs. Density Tradeoff

When packaged as a SSD, MLC/eMLC flash typically has a 4x density advantage (and an 8x cost advantage) as compared to SLC flash based SSDs. Though MLC has a significant write endurance disadvantage, it is often possible to trade off endurance for capacity; effectively increasing the endurance of MLC flash through system level overprovisioning. For example, if twice as much storage is provided than is actually needed or used (overprovisioning), the write endurance is effectively increased by that same factor, as writes will be spread out among twice the number of flash blocks, and each individual flash block will be subjected to only one-half of the program/erase cycles. When system level overprovisioning is combined with the use of eMLC flash, the effective endurance can actually exceed that of SLC flash.

Data Retention

Data retention is the ability of flash cells to store data. Data storage in flash relies on retaining stored charge levels. These charge levels will degrade over time (typically over many years), and this degradation becomes worse at high temperature extremes. The more bits that are stored in each flash cell, the more critical this degradation becomes. (This is the primary reason that TLC flash is only recommended for carefully selected applications) Current generation MLC flash has very sophisticated flash block and threshold management and error correction and use internal RAID methodologies that increasingly protect against this degradation. But nonetheless if the best possible long term data retention and/or write endurance is needed, especially at high temperatures, SLC flash may be the best choice.

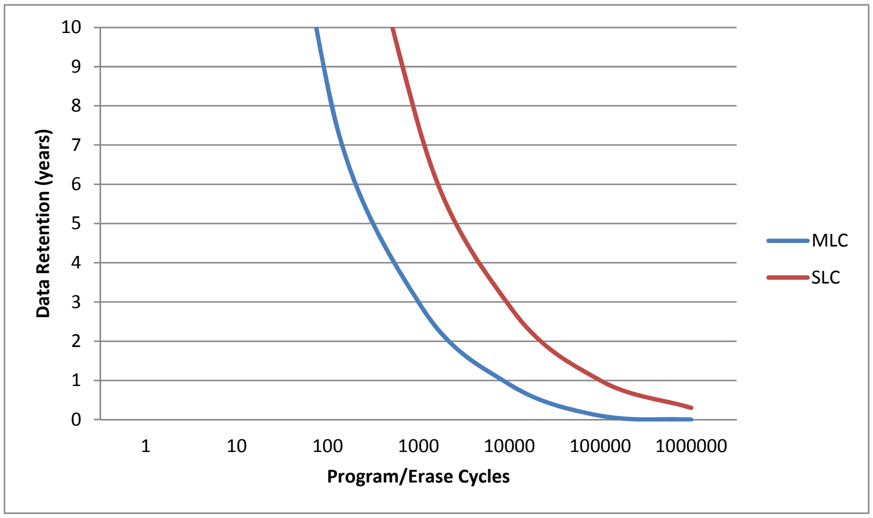
Endurance and data retention are tightly coupled. A SSDs data retention is best when it is brand new, and its retention ability steadily declines as the SSD is subjected to increasing program/erase cycles. It is in fact the SSD's reduced retention ability that generally defines the SSD endurance limit.

A key SSD specification is the Unrecoverable Bit Error Ratio (UBER). Typical UBERs for SSDs range from 10^{-15} for "client" SSDs to 10^{-16} for "enterprise" SSDs up to as high as 10^{-17} for "industrial" SSDs. The underlying "raw" bit error rate for the flash within the SSDs is much higher than the UBER. All SSDs employ intensive processing and error correction (ECC) functionality that is used in conjunction with reading data from flash blocks. As flash cells are subjected to increasing program/erase cycles, raw data reads from flash cells may degrade to the point that the ECC is not sufficient to allow data to recovered, resulting in an uncorrectable bit error. The endurance specification for an SSD is defined as the point at which, after a defined period of retention (typically 1 year) the SSD can no longer meet its specified UBER.

A typical relationship between endurance and data retention for representative MLC and SLC flash memory is shown in figure 1.

Note that this data is presented only to show an approximate relationship, and should not be viewed as precise data. The actual data retention period that will be realized varies greatly with the specific flash technology used, the programming and erase methods, the amount of error correction applied, and the operating and storage temperatures that the flash is subjected to.

Figure 1. Approximate Relationship between Endurance and Retention



The four most significant mechanisms that affect the ability of the flash cell to retain data are:

- Program Disturb
- Read Disturb
- Leakage
- Charge Trapping

Program disturb is a stress to cells that are in the same erase block as a cell that is being programmed. Most flash controllers will program pages sequentially within an erase block to minimize this effect.

Read disturb is caused by the voltage differential between a selected page that is being read and adjacent, unrelated pages. This can slightly stress the cells in these adjacent pages thus causing a small amount of charge to be transferred to the gate of an erased cell, essentially weakly programming the unrelated cell.

Leakage of the charge on the floating gate is a mechanism that can reduce the data retention time for a flash cell. The flash floating gates can lose electrons at a very slow rate, on the order of several electrons per month. This leakage rate is increased as the SSD is subjected to increase P/E cycles, and is increased at higher temperatures as well, resulting in reduced retention time. Cells are typically refreshed by the flash controller (by writing new data or refreshing of existing data) to mitigate the effects of leakage.

Charge trapping is the only one of the error mechanisms that can cause a permanent change to the cell that cannot be reversed by erasing or refreshing the cell. With every program or erase cycle, a small number of electrons can get trapped in the insulating oxide between the channel and the floating gate, causing a permanent shift in the voltage threshold and narrowing the gap between the erased and programmed states. Flash controllers mitigate this effect by adjusting programming and erase parameters and adjusting read thresholds dynamically with increased flash usage.

Retention, Endurance, and Storage Temperature

Power-off data retention time is also closely related to storage temperature; the higher the storage temperature, the shorter the data retention time. The write endurance specification for a SSD also implies a data retention duration. For example, an MLC SSD may specify an endurance of 3000 full drive writes, while maintaining data retention period of 1 year at a storage temperature of 40C. But when this SSD is brand new, the data retention time will be much longer (tens of years). On the other hand, if the SSD is stored power-off at a greatly elevated temperature, the data retention time will be reduced.

A Note About TLC Flash

Critical I/O uses TLC flash in a small set of our highest capacity storage products. Users should recognize that TLC flash has reduced write endurance and data retention as compared to MLC and SLC flash. TLC flash is most suitable for applications where a large volume of data is occasionally captured and must be stored for a short period of time prior to being offloaded to another storage medium.

An example of a suitable application might be an airborne sensor that flies occasional missions, with the SSD storage removed after each mission and taken to a ground facility for offload of data. The SSD storage is then erased and is available for the following day's mission. In such an application, TLC flash would remain usable for several years. Ultimately, many TLC flash based SSD applications should be considered a consumable commodity when used in a write-intensive environment.

4. SSD and Flash Controller Architectures for MLC Flash

SSDs are much more than raw flash memory. The controllers within SSDs are typically multi-core processor systems that manage eight or more banks of flash memory. While the most obvious function of the controller is to make raw flash look like a disk drive, SSD controllers also implement sophisticated flash management methods to overcome the inherent issues associated with using raw flash memory. In particular, the bit error rate (BER), the read/write performance, and the page/block oriented architecture of raw flash memory is unacceptable for use in most system applications, particularly so for raw MLC flash memory.

The consumer SSD market has driven MLC flash controller development technology in particular to a highly sophisticated level, and the embedded SSD user directly benefits from this technology development. Key areas of SSD and controller architecture include:

Error Correction - SSD controllers address the raw flash BER problem by storing additional error correcting codes with each data block, allowing multiple errors in a data block to be corrected upon reading the data. SSD controllers also typically dynamically tune internal flash read and write parameters to account for flash and data aging and temperature effects. MLC based SSDs in particular have very sophisticated error correction and read/write management implementations.

Multiple Flash Banks for Performance – SSDs address the flash performance problem through the use of multiple parallel banks of flash. As data is written to the SSD, the data is striped across these multiple banks, hiding the inherent flash memory write latency and aggregating the write performance of the flash banks.

Program and Erase Block Management - Flash memory within the SSD is organized as pages and blocks; a page (typically 4K to 16K) is the smallest unit of flash that can be written, while a block (typically 1MB and larger) is the smallest unit of flash that can be erased. Data can be written to flash on a block by block basis, but as the flash fills and existing data must be overwritten it becomes necessary to erase the flash prior to writes. As flash can only be erased on a full block basis this sometimes necessitates moving existing data occupying a partial flash block to a newly erased block in a process known as garbage collection, which may also result in write amplification.

Overprovisioning, Wear Leveling and Bad Block Management – With use, flash cells gradually wear out, to the point that they become unusable. SSD controllers implement wear-leveling in an attempt to ensure that all flash cells wear at a uniform rate. But even so, some flash cells will reach their end of life sooner than others, so SSDs maintain a supply of replacement flash blocks that can be swapped in to replace a defective block. The supply of spare flash blocks is known as over-provisioning, and can range from 7% to as much as 50% of the SSD flash capacity. Thus an SSD with a stated capacity of 500GB may contain as much as 1TB of raw flash. Overprovisioning also allows much more efficient management of flash block, providing higher and more consistent SSD performance.

Internal RAID – Many SSDs implement internal RAID like mechanisms where data is striped, with parity, across multiple flash blocks. In the event that a flash page should become completely unrecoverable due to excessive raw flash read errors, the data can still be reconstructed upon a read and relocated to a new flash block

5. What Type of Flash is the Best Fit for My Application?

The key application level factors to consider when selecting a specific type of flash listed below.

1) What is the capacity requirement?
2) What are the peak performance requirements?
3) How much data will be written? (*average* write rate x hours of operation)
4) How long data must be retained?

While each application will generally have different set of key requirements, many applications can be lumped together in several general categories. Four specific examples of typical systems are described below, with suggestions for the appropriate flash technology choices for each system. The system requirements and flash suggestions are also summarized in table 3.

Example 1: NAS File Server

File servers are generally read intensive, generally with relatively low read and write bandwidths. Continuous operation is the norm, with a need to retain data for long periods.

- Peak read rate: 200 MB/s
- Peak write rate: 100 MB/s
- Average write rate: 25 MB/s
- Capacity: 1.5 TB
- Operation: 10 hours per day

MLC or eMLC flash media are most cost effective choices for this application, as performance and capacity requirements are very modest, as are write rates. A typical SSD lifetime in this type of application using MLC flash media would be approximately 9 years, and if using eMLC media approximately 45 years.

Example 2: High Performance DAS storage

Generally a mix of reads/writes at moderate bandwidths, continuous operation, must retain data for long periods.

- Peak read rate: 200 MB/s
- Peak write rate: 200 MB/s
- Average write rate: 100 MB/s
- Capacity: 6 TB
- Operation: 24 hours per day

eMLC flash media is the most cost effective choices for this application, for this application, as the endurance requirement is still relatively low. A typical SSD lifetime in this type of application using eMLC media will be approximately 19 years.

Example 3: High Performance Data Recorder – Mission Oriented Operation

Generally very high bandwidth writes, short duty cycle operation, filling storage once per mission, offloading the data soon thereafter, with a limited data retention requirement.

- Peak read rate: n/a
- Peak write rate: 1 GB/s
- Average write rate: 1 GB/s
- Capacity: 6 TB
- Operation: 1 mission per day

MLC flash media is the most cost effective choice for this application, for this application, as the endurance requirement is still relatively low due to the SSD being filled sequentially only once per day. A typical SSD lifetime in this type of application using MLC flash media would be approximately 11 years, increasing to 55 years using eMLC flash.

Example 4: High Performance Data Recorder – Continuous Operation

Continuous high bandwidth writes, 100% write cycle operation when operational, continuously capturing data and overwriting old data, with a limited data retention requirement.

- Peak read rate: n/a
- Peak write rate: 500 MB/s,
- Average write rate: 500 MB/s
- Capacity: 6 TB
- Operation: 12 hours/day

eMLC or SLC flash media are the best choices for this application, for this application, as the endurance requirement is high due to the continuous write environment. A typical SSD lifetime in this type of application using eMLC flash media would be approximately 15 years, or approximately 46 years using SLC media.

Table 3. Summary of Storage Requirements for the Example Systems

Example Number	Sustained Write Rate (MB/s)	Size (GB)	Access Pattern	Flash Type	Size Weight Power	Assumed Flash Endurance	Usage Hrs/Day	Usage Fills/Day	Est SSD Lifetime Years
1	25	1,500	Random	MLC	1x	2,000	10		9
1	25	1,500	Random	eMLC	1x	10,000	10		46
2	100	6,000	Random	MLC	1x	2,000	24		4
2	100	6,000	Random	eMLC	1x	10,000	24		19
3	1,000	6,000	Sequential	MLC	1x	4,000		1	11
3	1,000	6,000	Sequential	eMLC	1x	20,000		1	55
4	500	6,000	Sequential	eMLC	1x	20,000	12		15
4	500	6,000	Sequential	SLC	4x	60,000	12		46