

Reliability "Facts of Life" for Hard, Flash and DRAM Drives in High Performance Storage Systems

White Paper
August 2009

Abstract

This paper is intended for IT executives and managers who are involved in development of **highly reliable high-performance storage systems**. In Developing high performance storage architectures you need to consider the performance and reliability characteristics of each storage type. The use of HDDs, flash based SSDs, and DRAM based SSDs all should be considered. This paper addresses the reliability characteristics and considerations of each of these types of storage as well as the system level implications to consider in using them. For the highest performance environments, the use of DRAM based SSDs (like the jet.io RamFlash drive) will provide the most reliable solutions as well as the highest performance and lowest cost.

Table of Contents

1. Introduction: Real World Reliability in High Performance Storage Systems (HDD, Flash SSD, DRAM SSD)	3
2. HDD failure rates.....	4
2.1. Real-World HDD Failure Rate Estimates	6
2.2. Performance affect on HDD reliability	6
3. Flash Based Solid State Drive (F-SSD) failure rates.....	6
3.1. Flash Durability Basics.....	7
3.2. Modern Flash SSDs.....	8
3.3. Estimating Flash SSD Drive level Reliability	9
3.4. Cutting through Flash Marketing Information	11
4. RAM Based R-SSD failure rates	12
5. System Level Failure Rates	13
6. Summary.....	15
i. Peak to Average Workload Calculation:.....	16
ii. Solid State Drive Types and Acronyms.....	16
iii. References.....	18

1. Introduction: Real World Reliability in High Performance Storage Systems (HDD, Flash SSD, DRAM SSD)

In architecting the optimum storage system there are many tradeoffs to make including system level decisions and selection of the storage components. Today the best solutions include using a tiered approach with the optimum mix of traditional spinning HDDs, flash based Solid-State Drives (F-SSDs), and RAM based Solid-State Drives (R-SSDs) as well as tradeoffs of array types, caching units, software, and quantity of servers. All of which tie together in an attempt to achieve a highly reliable system with the highest performance, lowest cost, and lowest ongoing support and maintenance cost. This paper will address the issue of Reliability as part of the decision making process.

Reliability data reported by HDD and SSD drive manufacturers (**MTTF/MTTR rates or drive life estimates**) are **nearly meaningless** for determining failure rates and failure modes **in high performance storage systems**. In order to understand the true reliability of your system you must look at real-world failure rates taking into account the age of the product and the operating conditions including the characteristics of the storage usage (operations/second, randomness).

The Mean-Time-Between-Failure (MTBF) of HDDs, as specified in datasheets is typically in the 1M hour range which would equate to 114 years mean time to failure. Given that hard drives have only been in existence for 52 years you might wonder how this type of specification could be given.

Most IT experts understand that these values are “theoretical” MTBF values and are not intended to be an indicator of the true expected life of each hard drive, but merely as a value used (assuming a large population of drives) to determine an annual replacement rate (ARR) or expected annual failure rate as a percentage of drives deployed as follows:

$8760/\text{MTBF}=\text{ARR}\%$ (where 8760 is hours in a year)

As an example typical HDD quoted MTBF values of 1,000,000 hours, suggest a nominal annual failure rate of 0.88%. Flash based SSDs typically quote MTBF values in the 5,000,000 range indicating that their ARR would be 0.17%. The real world failure rates in high performance storage environments will be much higher especially as the drives age and as they are used in intense high performance applications since both of these drive types (**HDD and flash based SSD**) have **well known wear-affected failure modes**.

Given that these MTBF estimates (even if correct) were developed using accelerated life testing, you might wonder how they were accelerated and how your actual environment and usage patterns might affect that life as well.

What you really need to know:

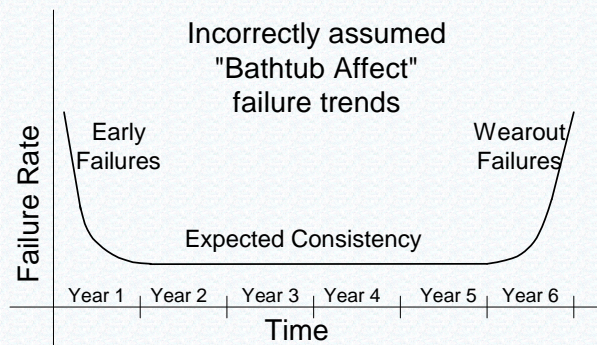
- 1) In my environment (high performance storage as covered in this white paper), what is the failure rate of each device taking into account length of time in service and usage characteristics.
- 2) For my application, taking into account the number of devices required and the type of storage configuration, what will be the failure rate across the entire solution and what is the probability of data loss in the storage array's based on realistic reliability data

This white paper will address these questions for traditional spinning HDDs, F-SSDs, and R-SSDs. We will also address how this information could be used to evaluate failure rates in a complete solution example.

2. HDD failure rates

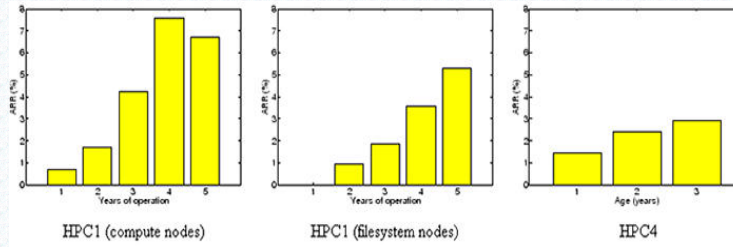
Spinning hard drive MTBF rates quoted in data sheets are not a reasonable indicator for annual replacement rates. This is true even for the average IT environment let alone the high performance storage environment with very high rates of random access sending the drives into constant motion.

In reality, as shown in a number of large scale HDD installation studies, the ARR is much higher than indicated by quoted MTBF values and it increases significantly every year. In fact in the past, it was assumed that HDDs would have a “bathtub” affect where it was assumed that there would be a higher failure rate due to manufacturing issues in year one, steady failure rates in years 2-5 and then higher failure rates after year 5 as follows (1).



In reality, the failure rates have been found to be more directly tied to the age of the drive. The following results from a real world HDD failure analysis (Reference 1) show what you can really expect from an average HDD deployment (taken from three sets of data in large disk deployments).

Real World Reliability in High Performance Storage



The June 2009 issue of **Storage Magazine** gave a good summary of this study (1), summarizing as follows:

“Unfortunately, manufacture MTBF numbers don’t reliably reflect real-world MTBFs. The Computer Science department at Carnegie Mellon University in Pittsburgh ran stress tests of 100,000 Fibre Channel, SAS, and SATA hard disk drives. Their published testing results determined that a typical drive has a realistic MTBF of approximately six years or 52,560 hours. Using Carnegie Mellon’s MTBF numbers a storage system with 240 HDDs can expect a drive failure approximately every nine to 10 days (approximately 40 HDDs per year or an **annual replacement rate of 16.67%**).”

The subject of actual drive failure rates is typically not addressed by drive manufacturers. In response to this article on real world HDD failure rates, most drive vendors declined to be interviewed. A Seagate Technology spokesperson replied:

"..... "It is important to not only understand the kind of drive being used, but the system or environment in which it was placed and its workload."

Indeed, this mention of workload is one of the issues that is not typically considered in developing a storage system. Without taking into account real world failure rates based on age of the drive, and workload, the resulting system design may have a high risk of data loss.

Ashish Nadkarni, a principal consultant at GlassHouse Technologies Inc., a storage services provider in Framingham, Mass., said “he isn’t surprised by the comparatively high replacement rates because of the difference between the “clean room” environment in which vendors test and the heat, dust, noise or vibrations in an actual data center”.

This accelerated life testing likely utilizes high rates of random drive access to stress the drives mechanical systems but a high performance computing environment will have the same effect, therefore shortening the life of the drives below that of lighter IT environments.

2.1. Real-World HDD Failure Rate Estimates

Taking into account the workload as suggested by Seagate, the additional wear and tear on the mechanical systems of spinning HDDs in high IOPS environments will affect failure rates especially in the later years of the drives life. This data could be extrapolated to provide a more realistic Annual Replacement Rate based on age of the drives and degree of random Input Output Operations Per Second (IOPS) as follows:

HDD Real World Annual Failure Rates at random IOPS usage rates

IOPS (peak)*	Year 1 ARR	Year 2 ARR	Year 3 ARR	Year 4 ARR	Year 5 ARR
25	1%	2%	3%	4.5%	6%
100	1.2%	2.6%	4.1%	6.5%	9%%
300	2%	6%	9%	16%	20%

2.2. Performance affect on HDD reliability

The other aspect of HDD use in **high performance applications** is that you typically **require large numbers of these drives to meet the high IOPS requirements** as well as the array controllers and drive chassis components required to house them. This aspect of the system design **will multiply the failure rates** overall, reducing the system reliability and risk of data loss.

As an example, in order to support 60,000 IOPS with HDD spinning drives, you would need at least 200 active hard drives plus the redundancy drives in RAID configurations in order to provide overall reliability of the system. If real world failure rates are not taken into account there could be catastrophic loss of data due to a 2nd drive failure in a RAID 5 array during the long drive re-build for an initial failure. Use of RAID 10 or RAID 6 configurations is recommended in these environments. And of course the use of tiered storage to reduce the high access rates on the mechanical HDD storage is another way to improve reliability of these systems.

This issue is mitigated with solid state drives that can achieve 5,000 IOPS (F-SSD) or 40,000 IOPS (R-SSD) in a single drive. System designs using solid state solutions for the high IOPS traffic will have considerably higher reliability.

3. Flash Based Solid State Drive (F-SSD) failure rates

F-SSDs have no moving mechanical parts like spinning HDD drives, but F-SSDs have inherent durability issues based on wear out due to write/erase cycles and data-retention limitations. These are similar and in some cases worse that the wear affects on HDDs. Because of these factors, the F-SSD quoted MTBF numbers and product life claims are not valid indicators of failure rates for these products when used in high performance storage environment. You must be careful in reading the specifications of F-SSD drives to know what caveats and conditions are assumed and to take into account the utilization rates of your applications. In certain applications, these drives could have a much higher failure rate than an HDD, but

the actual storage application characteristics must be taken into account to have a meaningful discussion of reliability.

3.1. Flash Durability Basics

Flash technology has two key constraints in terms of reliability which are maximum write/erase cycles per block, and data retention. These will vary depending on the type of flash used, and the overall drive reliability can be enhanced somewhat through complex controller designs and over-provisioning.

Data retention is the amount of time that the flash cell data will remain programmed and is generally inversely related to the number of write erase cycles.. Drives with minimal write/erase cycles might have a 10 year retention capability, whereas drives operated at near maximum write/erase capacity might have their retention capability reduced to 1 year.

The key durability constraint for flash is related to **endurance** or limitations of write/erase cycles. Single level cell (SLC) and Multi level cell (MLC) are the two types of flash technology used and these have very different durability characteristics. Generally **SLC** flash will support up to **100,000 write/erase cycles per cell** and **MLC** will support around **10,000 write erase cycles per cell**.

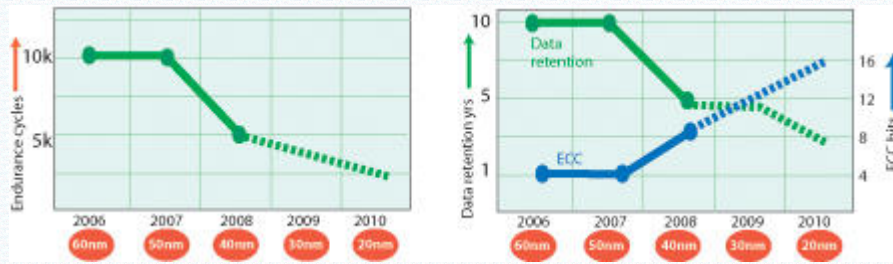
To put this into perspective, without complex controllers (changing address to cell relationship), writing to the same location 100 times per second would kill that cell of the flash in around 15 minutes even for the higher endurance SLC flash.

Another issue with flash is that it cannot be written to a byte at a time. Flash is organized in large blocks which must be erased before you can write to them. A typical design would have 512 Kbyte Blocks which are organized into 4 Kbyte pages (Note block size is increasing as flash size increases). Once a block has been written to, you must perform a read/erase/write cycle for any write and you can only erase whole blocks. This means that even if you only wanted for example to do a 512 byte write, you would actually need to erase and write the entire 512k block (**1,000:1 write amplification factor**). This “write amplification” issue can have a serious effect on real durability for applications with high volumes of smaller random write operations

To put this into perspective, a simple 32GByte flash SSD design that was accessed with 5,000 512B random writes per second (2.5MBps of load) would have the same level of wear on the cells as if it was being written to at a 2.5GBps rate. This is due to the amplification factor causing each 512 byte write to actually perform a 512kbyte read, erase, and re-write operation. Of course modern flash controllers can mitigate this affect somewhat.

As detailed in a reliability white paper by Samsung (5), as flash density increases, manufacturers are using smaller physical sizes for each cell which will contribute to

lower data retention and endurance capability. As summarized in the graphic below which is taken from an article by Samsung, their prediction for MLC endurance with newer larger capacity flash chips will be even lower than current MLC flash.



Flash devices will also have a finite error rate primarily from Soft errors in memory which are non-permanent errors (unless they occur during a write operation) which are not related to bad memory locations. Soft errors can be caused by poor power regulation, alpha particles or cosmic rays. These are a particular issue in MLC based flash devices because each memory cell has to discriminate between multiple voltage levels (at least 4 levels instead of two for SLC). This concern in addition to the inherent cell wear-out issues mean that flash designs must implement substantial error correction mechanisms to mitigate the effects of bit errors.

3.2. Modern Flash SSDs

The good news for F-SSDs is that there are techniques that can be used to provide some improvement of these reliability issues. Modern F-SSDs use complex controller designs in order to mitigate some of these durability issues and to enhance the write speed of the inherently slow flash chips. ***These algorithms often tradeoff drive life with performance and the details of the algorithms need to be taken into account when calculating drive life.*** These algorithm's also do not work the same for every traffic type and they may cause unexpected affects on performance over time (typically new/fresh flash drives will have very high performance, but over time the performance will drop and drives that are accessed heavily will have widely varying and dramatically lower performance as detailed by numerous tests (see references 6, 7, and 8). Some vendors have provided firmware updates to improve on these performance issues, but the algorithm changes will need to be re-evaluated in terms of their affects on product life.

These flash controllers provide a "Wear Leveling" mechanism to allow the drive to age evenly across all blocks. In addition there are some techniques using over-provisioning which can reduce the write-multiplication affect. Some of the drives do a better job than others with these algorithms such that some will have better durability and some will have better and more consistent performance. **One issue is that there is no well defined standard for defining how the specific controller design will affect durability in your application.**

The use of RAM cache in F-SSDs also can provide improvements in performance and reliability. It has been suggested by StorageSearch.com that F-SSDs be put into categories to define the amount of RAM cache which it termed Fat-SSD and Skinny-SSD. The Fat-flash SSD drives typically have somewhat better performance and durability, but must also have battery and backup capability for the cache similar to a R-SSD.

Unfortunately, the key to understanding the true reliability of a flash based F-SSD is to delve into the details of flash type and make, and controller algorithms which the manufacturers are not likely to provide. As flash drives are pushed to provide more density, the use of smaller cell geometry and more bits per cell (MLC) will drive the Real World failure rates higher especially for high performance storage applications.

For applications with less intense random write requirements and more heavily read driven requirements, F-SSDs could provide a reliable option with improved performance and failure rates compared to spinning HDDs.

3.3. Estimating Flash SSD Drive level Reliability

As mentioned above developing the overall reliability characteristics of F-SSDs requires evaluation of a number of things including the type of flash chip used, the amount of over-provisioning within the drive, and the details of the drive design especially the wear leveling algorithms. The following tables provide an estimate of real-world reliability in terms of failure rates for drives based on write operations per second and age of drive.

In order to use more conservative usage rates and address the best possible F-SSD designs, the following failure rate estimates assume that

- The Average workload is assumed to be 25% of the peak workload (see section i at the end of this paper regarding workload assumptions)
- An ideal flash controller is assumed to provide perfect wear leveling such that no cells are effected more than others
- It is also assumed that the F-SSD uses substantial over-provisioning (extra flash) and an effective algorithm such that a block erase is only done once for every 5 random write operations

These assumptions only apply to an ideal F-SSD design but still result in a solution that will wear out quickly for the 5000 peak write OPS case (less than a couple of months for an MLC drive, and about 1.5years for SLC).

Real World Annual Failure Rates for a 32G MLC F-SSD

W-OPS (peak) *	Year 1	year 2	year 3	year 4	year 5
100	0.5%	0.6%	1.0%	3.0%	5.0%
1000	25.0%	70.0%	100.0%	100.0%	100.0%
5000	60.0%	100.0%	100.0%	100.0%	100.0%

Real World Annual Failure Rates for a 32G SLC F-SSD

W-OPS (peak) *	Year 1	year 2	year 3	year 4	year 5
100	0.5%	0.3%	0.3%	0.3%	0.3%
1000	0.5%	5.0%	10.0%	20.0%	40.0%
5000	0.5%	25.0%	80.0%	100.0%	100.0%

* W-OPS is Write Operations Per Second

These calculations are also done using the more realistic maximum of 5,000 write OPS for the F-SSD versus the higher performance claimed in many datasheets. **If we used the claimed write OPS performance of some flash drives as the peak in these charts, the drives would last for days or weeks and not years.** These claims are usually related to an ideal test environment with a fresh unused drive, and are not realistic for long term performance.

These extrapolations of failure rates indicate that flash can be very reliable under low write OPS conditions, but that well understood flash endurance issues will cause failures over time when the applications are write intensive.

While the best SLC flash drives can achieve around 5,000 sustained W-OPS, using them in an application that requires this level of peak write workload (even assuming the average is 25% of this peak) would result in an unacceptable product life (more than a year at these write workloads would be very risky).. For this reason, **even SLC based SF-SSDs should only be considered in applications where the peak write workload is about 1,000 Write operations per second which is still about 4x better than current spinning HDDs. MLC based MF-SSD drives should only be considered where the write workloads are about the same as can be achieved by HDDs.**

You must also keep in mind that these calculations assume an ideal flash controller design with substantial over-provisioning. You will need to evaluate each drives characteristics in addition to your workload in order to develop a reliable system design.

3.4. Cutting through Flash Marketing Information

Every manufacturer uses marketing adjustments to help make a product seem better than it is through the use of a number of techniques. These include omission of specifications that may cause customer concern, and making outrageous claims that are only true in unrealistically constrained circumstances. In order to understand the true reliability of these products we must request additional details and extrapolate what we do know into usable values.

Flash based SSDs have come a long ways, but their product specs can be particularly misleading. It is well understood that the best of these has great read performance and is very reliable when lightly utilized. **The missing specs are real sustained random write performance (full drives during heavy use), and real product life expectations based on differing write work loads.**

For example, an 80 GB PCIe flash based SSD product from a top provider quotes product life of 25 years but caveats that this is assuming 5TB of writes per day and not defining characteristics of those writes (512 bytes or 512kbyte per write makes a big difference due to the write amplification factor described above). Using the same ratio of daily use to peak hour as used in the reliability tables above, this indicates they are quoting product life on a real-world **peak-hour** write workload of somewhere between **0.5 write operations per second** and 450 write operations per second (likely somewhere between).

Conversion of product life specs to equivalent peak hour write operations per second			
Assuming 512k record size (highest product life)		Assuming 512B record size (highest write OPS)	
5.00E+09	5TB of write per day in drive life specs	5.00E+09	5TB of write per day in drive life specs
24	Hours per day	24	Hours per day
3,600	Seconds per hour	3,600	Seconds per hour
512,000	Record size for writes (best case for drive life)	512	Record size for writes (worse case for drive life)
0.11	Write operations per second average for day	113.03	Write operations per second average for day
4	Peak to average	4	Peak to average
0.45	Write operations per second peak hour	452.11	Write operations per second peak hour

In trying to understand this highly touted flash drives product life, it isn't clear which of these was intended as the product life specification, but even using the higher peak workload, this is about what you can get from a spinning HDD drive.

This same drive vendor quoted a mix of read/write that equates to 22,500 write operations per second. However in real world benchmarks it has been shown to support approximately 9,000 operations per second when under heavy load and extended use.

Based on even the more realistic write workload of 9,000 write OPS for peak hour (or 2,250 write OPS average) and the best case from above, **you could extrapolate that this flash based SSD drive would have a best case product life of 1.25 years.** if you could operate this drive **at its quoted 22,500 write OPS workload you could extrapolate that it would only have a product life of 6 months**, If the product life specification were based on larger write records sizes, then this extrapolation would lead to much shorter product life estimates.

4. RAM Based R-SSD failure rates

RAM based R-SSDs like the flash based F-SSDs have no moving parts but in contrast do not suffer from the wear-out failure modes of flash under high utilization. These devices will therefore have extremely low and consistent failure rates over the full 5 year product lifecycle. Typical calculated failure rates for these types of devices are around 6M hours, but a more realistic failure rate will likely be around 0.5%. Due to the use of RAM for storage, **the rate of failure for R-SSDs will not be affected by age or work load** like the spinning HDDs or the flash based F-SSDs. This results in an expected reliability as follows including the assumption of a much higher usable write operations per second:

Real World Reliability (ARR) for RAM based SSD (32G)

W-OPS (peak) *	Year 1	year 2	year 3	year 4	year 5
100	0.5%	0.5%	0.5%	0.5%	0.5%
1000	0.5%	0.5%	0.5%	0.5%	0.5%
5000	0.5%	0.5%	0.5%	0.5%	0.5%
20000	0.5%	0.5%	0.5%	0.5%	0.5%

* W-OPS is Write Operations Per Second

RAM devices can also have a finite error rate from Soft errors but these will be at a much lower rate than flash or HDD magnetic storage media. Soft errors in memory are non-permanent errors which are not related to bad memory locations. Soft errors can be caused by poor power regulation, alpha particles or cosmic rays. While the rate of these types of errors in RAM based SSD is much lower than with HDD or flash based SSD, the use of error correction should still be included as part of the drive design and should improve this error rate effectively to zero.

With RAM based SSDs you also have to consider the non-volatility of RAM and the use of battery and backup (non-volatile storage) as part of the reliability equation. For backup storage these drives could use flash and this is where flash is very reliable since it is only written to in a controlled manner for backup of the main storage.

Batteries in the RAM based SSD also need to be considered in terms of reliability. In considering battery failure rates, the operating conditions must again be taken into account. In enterprise storage environments, these batteries will be in controlled IT data centers and will not be constantly charge cycled (like a smart-phone for instance) since the units should be constantly operational except for the unlikely case of power loss or for special maintenance reconfiguration. In these conditions with minimal charge cycling and assuming state of the art Lithium-Ion battery designs, the failure rates of these batteries should be very low and consistent through the designed product life (5 years). In addition, it must be taken into account that battery function will not affect active storage operation and that these SSD designs should provide a mechanism to monitor the batteries to detect

conditions that could signal potential failure. This could allow the RAM based SSD's to be serviced for battery replacement during maintenance windows and allow the drive life to be essentially unaffected by battery failures.

It is debatable whether the use of backup drives and batteries substantially affects the reliability of RAM based SSDs, but in either case it must be remembered that this affect would be very small in comparison to the affect that flash wear-out has on the failure rate of Flash based SSD's used in high write OPS workloads. In addition, this affect must also be taken into account for the Fat F-SSDs which will also have battery backup for their RAM cache.

*An example RAM based SSD from Density Dynamics, **the jet.io drive, uses a RamFlash innovation which leverages the resiliency and ultra-fast access of DRAM based memory as primary storage with the non-volatile nature of flash memory to deliver breakthrough performance with ultimate durability and data protection.** This design uses a number of architectural design features to further minimize the potential for failures below that of DIMM based designs by using low voltage DDR2 memory running at low clock rates and utilizing proprietary power reduction techniques to minimize temperature in the drive. Any soft errors are mitigated through the use of ECC error correction. This design utilizes a Flash backup drive and high reliability Lithium-Ion battery that is fully monitored to provide advance warning upon detection of conditions indicating potential battery failure.*

The real advantage of the RAM based SSD is that it has **the reliability of a solid state device regardless of the work load applied.** RAM based R-SSDs can also achieve higher write operations-per-second workloads such that you can use less drives to achieve the same total solution compared to flash based F-SSDs resulting in even higher reliability per workload.

5. System Level Failure Rates

It is important to address not just the individual storage drive failure rates, but to also consider the reliability of the total storage solution. For high performance storage applications, you must compare solutions based on workload, performance requirement, and desired reliability when developing the overall system design. This section will address the differences in individual drive failure rates using simple mirrored array designs and taking into account the different number of drives required for each of the types of storage.

For operations that intend to utilize RAID 5 configurations, a more thorough evaluation should be done to include the increased risk of data loss using RAID 5 due to increased probability of a second drive failure while rebuilding an array following an initial drive failure.

A typical high performance storage requirement will be used to characterize this system level reliability difference between the three drive types. This requirement is:

- 60,000 IOPS of random load with a 50% read/write factor
- 128 GB data set

To investigate the total system failure rates, we will use a simple calculation for system level drive failure rate as follows:

$$MTBF = 1 / \sum(\lambda_1, \lambda_2, \lambda_3, \dots \lambda_N)$$

Where λ_N is the annual failure rate of each component and MTBF is the mean time between drive failures in years.

HDD

For this example requirement, the spinning HDD solution must meet the high IOPS requirement by spreading the dataset across a large number of drives since each drive can only achieve around 300 IOPS. This means a total of 200 active hard drives are used to address this requirement. In addition, these drives will likely utilize multiple array controllers and a lot of drive chassis. In an effort to remain reliable, these drives will also be protected through the use of mirroring or another protection level such as RAID5 or RAID6. For simplicity we will assume a mirrored RAID 1 configuration, which doubles the requirement to 400 drives.

If we simplify the above spinning drive example to just look at the drives (ignoring the failure rates of the large number of chassis, Array controllers, and Caching units) and assume a failure rate of 16%, the rate of drive failures for this total solution will mean that **a drive will fail every 6 days**. In determining the likelihood of total data loss, you must also consider the probability of a second drive failure while the array is re-built after a failure. For RAID 1, this is still fairly low, but is a concern. For RAID 5 configurations you would use less total drives, but the risk of having a second drive fail in an array during array rebuild is very significant due to the long re-build times for this RAID configuration.

Flash based F-SSD

As mentioned in section 3.3, while some F-SSDs can attain sustained performance of around 5,000 random write OPS, even SLC based F-SSDs would not provide a reliable solution at this kind of peak hour workload. For this specific application of 60,000 IOPS one option is to spread the write load across a larger number of drives (as is done with low IOPS spinning HDDs) in order to bring the long term failure rates down to a more reasonable level. In this case if we assume we only loaded each F-SSD with 1000 W-OPS, you would need 30 active SF-SSDs. In a mirrored RAID1 configuration that would equate to **60 32G SLC based F-SSD drives**. Even with SLC based F-SSDs, the failure rates in later years at this workload are still very significant such that **in year 4 of use the failure rate for just the drives in this configuration would be a failure every 30 days**. This configuration will allow the use of well designed SLC flash drives in high performance storage, **but the number**

of drives required for the workload must be a consideration when evaluating the cost of the total solution.

If MLC based MF-SSDs were used, you would need to limit the W-OPS to a much lower level to provide reasonable levels of reliability. An example might be to use 300 W-OPS on each drive which would mean a requirement for $30,000/300 = 100$ MLC flash drives. If you include a mirrored RAID1 configuration **the MLC solutions would require 200 flash drives.** This solution is similar to that of the spinning HDD drives but at a much higher cost. Really the MLC solution does not make sense at all for high write workload high performance storage applications.

RAM based R-SSD

Using RAM based R-SSD, you can assume at least 20,000 write OPS per drive with no long term reliability concerns. Four drives are required to handle the 128GB dataset and a **total of 8 drives** with mirrored RAID1. **In this case, the System level MTBF calculation indicates that this 8 drive configuration would last well beyond the 5 year product life with no failures.**

6. Summary

In summary, HDD, flash based F-SSD, and RAM based R-SSD are all reasonably reliable for light workloads, but for high performance storage environments you need to take into account the workload and years in service to calculate the real failure rates and risk of losing mission critical data. In developing the best overall strategies for your storage taking into account performance requirements and reliability concerns, the following should be considered.

- The use of tiered architectures will help make the best use of these three storage technologies.
- RAM based R-SSD will provide a far better solution for the highest performance applications especially those with high write workloads
- SLC flash based F-SSD (SF-SSD) will provide good solutions for heavy read dominant workloads with lighter write workloads
- MLC based F-SSD (MF-SSD) are suitable for read-only workloads and very light write workloads
- Spinning HDDs can provide good tradeoffs of price/performance in many cases and are very good where sequential access is the main access mode
- Capacity HDDs are still the best choice for less accessed storage of large data sets.

i. Peak to Average Workload Calculation:

The following calculation shows one assumption for the daily average of write operations per second based on a peak hour requirement. For a given application, analysis will yield a higher or lower percentage of peak to average than this example. A very conservative value of 25% was used in this paper. Many high performance storage environments will yield nearly 100% of peak levels 24 hours per day which will exacerbate the issues discussed in this paper regarding wear related failures in HDDs and flash based F-SSDs.

Daily Application Work Load		
Average write Operations per second		
Peak Hour	100%	5,000
Medium Load (6 hours)	80%	4,000
Light Load (5 hours)	50%	2,000
Off Hour (12 hours)	20%	400
Average Hour Workload		1,825
Percentage Average to Peak		37%

ii. Solid State Drive Types and Acronyms

As you can see from this white paper, there are significant differences between the different categories of Solid State Drives and even differences within a category depending on the details of which chips are used, and how the controllers are designed considering tradeoffs of performance and durability.

The following categories are proposed and used in this white paper, although drives falling into these broad categories will still vary significantly due to design differences and other characteristics.

SSD:	the generic category for solid state storage often used for the largest category of flash based SSDs.
F-SSD:	Flash based SSD of either MLC or SLC type. This category covers a wide range of SSD products including low performance MLC based drives and higher performance SLC based drives. The amount of RAM cache used in these drives will also affect performance and durability

R-SSD:	RAM based SSDs. These drives are the highest performance SSD option and will provide the lowest latency and highest reliably sustainable write performance (typically 10x to 20x better than F-SSD). RAM storage technology does not suffer from cell wear out and therefore these drives can be operated at their highest write workloads indefinitely. RAM is a volatile storage technology, so R-SSD must be used in a way to provide power backup and with software to backup the data on the R-SSD in the event of power loss.
MF-SSD:	MLC flash (multi-Level Cell) based SSDs. This category has the widest range of performance characteristics with some drives similar in performance to spinning hard disk drives. Most MF-SSDs will have better Random access latency than HDDs, have very low power draw, and provide a solid state design with no moving parts. These drives can be a good option for storage applications that are read intensive. Due to cell wear-out characteristics (allowing 10,000 erase/write cycles per cell) these drives should only be used where sustained write workloads do not exceed 300 to 500 write OPS.
SF-SSD:	SLC flash (single-level Cell) based SSDs. These are a higher performance version F-SSD, using SLC flash parts which have better cell wear-out characteristics (100,000 erase/write cycles per cell). Many of these SF-SSDs are designed as "enterprise grade" SSDs and are more suitable for mixed workloads. Most of these drives offer high random read OPS, and some are capable of fairly high instantaneous write OPS. For reliable system design, these SLC based drives should only be used for sustained peak write workloads below 1,000 random write operations per second.
Fat F-SSD:	Flash based F-SSDs which have a large RAM cache. This terminology was proposed at http://www.storagesearch.com/ram-in-flash-ssd.html . These drives typically have somewhat better performance and durability, but must also have battery and backup capability for the cache similar to the RF-SSD. Do not confuse "Fat F-SSDs" with R-SSDs or RF-SSDs which have RAM for the entire storage space in the drive. Fat F-SSDs RAM cache will only be a small fraction of the total storage capacity
Skinny F-SSD:	The opposite of Fat-SSDs. These drives have no RAM cache which will have a significant affect on the performance and durability of the drive.
RF-SSD:	RamFlash SSD. This is the Density Dynamics jet.io drive innovation which leverages the resiliency and ultra-fast access of DRAM based memory as primary storage with the non-volatile nature of flash memory to deliver breakthrough performance with ultimate durability and data protection. This is not to be confused with what has been termed Fat Flash SSDs (described above) which use an amount of DRAM for cache that is a fraction of the total drive capacity. The RamFlash technology uses DRAM for all primary storage capacity and only uses the flash capacity as a backup copy which is written to in a controlled manner to avoid any flash write durability or performance issues.

iii. References

- 1) 5th USENIX Conference on File and Storage Technologies, Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?
Bianca Schroeder Garth A. Gibson Computer Science Department Carnegie Mellon University
- 2) <http://techreport.com/articles.x/15931>
- 3) http://www.siliconsystems.com/technology/wp_NAND_Evolution.aspx
- 4) <http://www.behardware.com/articles/731-2/ssd-product-review-intel-ocz-samsung-silicon-power-supertalent.html>
- 5) <http://www.ecnmag.com/Web-Exclusive-Maximizing-MLC-NAND-Flash-Reliability.aspx?menuid=578>
- 6) <http://www.engadget.com/2009/02/19/intel-x25-m-ssds-slowng-down-with-extensive-use/>
- 7) <http://forum.ssdworld.ch/viewtopic.php?f=1&t=59>
- 8) <http://forums.storagereview.net/index.php?showtopic=27190>