Lenovo

eXFlash and S3700 Enterprise SSDs: Technology Overview

Last Update: July 2013

Describes the 1.8-inch SSD offerings for System x servers

Covers the Intel S3700 and S3500 Enterprise SSDs

Explains the technology behind these solid-state drives

Explains typical usage scenarios for these SSDs

Ilya Krutov

Abstract

Ensuring that business-critical data is available when needed is an ever-growing need in IT. Your systems must store massive amounts of data quickly and retrieve it efficiently. Simultaneously, you must use new technologies that can improve efficiency and take advantage of these technologies within limited budgets.

One measure of growing efficiency in recent years is CPU processing power, which far exceeds growth in disk input/output (I/O). For this reason, disk I/O is often the reason for bottlenecks in high-performance applications.

The combination of the eXFlash™ unit and the S3700 MLC Enterprise or S3500 MLC Enterprise Value 1.8-inch solid-state drives (SSDs) can help to close the performance gap between CPU compute power and storage I/O, providing cost-optimized performance for different types of storage-intensive applications.

This paper discusses server performance imbalance that can be found in typical application environments, and how to address this issue with the S3700 or S3500 SSD technology that can be utilized in the eXFlash solution to provide required levels of performance and availability for the storage-intensive applications.

At Lenovo® Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Contents

Executive summary

Currently, the processor, memory, and I/O subsystem are well balanced and virtually not considered as performance bottlenecks in the majority of systems. The major source of performance issues is related to the storage I/O activity because of the speed of traditional hard disk drive (HDD)-based storage systems that still does not match the processing capabilities of the servers.

This can lead to a situation when a powerful processor sits idle waiting for the storage I/O requests to complete, therefore wasting its time, which negatively affects user productivity, extends the return on investment (ROI) time frame, and increases overall total cost of ownership (TCO).

The eXFlash offering can help to address the issues described above by combining innovative high-density design of the drive cages and the performance-optimized storage controllers with the reliable high-speed solid-state drive technology. This allows eXFlash to significantly decrease storage I/O response time to match the processing power of the server CPUs without a need to deploy high-performance external storage system.

eXFlash, combined with the S3700 MLC Enterprise SSDs, can help to achieve:

- Cost-optimized performance and efficiency for read-intensive and write-intensive enterprise applications, such as databases, corporate email and collaboration, and actively connected users (Web 2.0)
- Over 200,000 input/output operations per second (IOPS) (RAID-5, 4 KB blocks, online transaction processing (OLTP) workload) for each eXFlash unit
- Sustainability to heavy write operations with the endurance that allows a single SSD to be fully rewritten up to ten times a day during the five-year lifetime expectancy
- Higher reliability and availability of internal storage due to absence of moving parts
- Shorten ROI time frame and decrease overall TCO with the efficient usage of server resources and lower power, cooling, and management costs

eXFlash, combined with the S3500 MLC Enterprise Value SSDs, can help to achieve:

- Cost-optimized performance and efficiency for read-intensive enterprise applications, such as web serving, content delivery, streaming video, and big data
- Up to 4 GBps of sustained read throughput for each eXFlash unit
- Higher reliability and availability of internal storage due to absence of moving parts
- Shorten ROI time frame and decrease overall TCO with the efficient utilization of server resources and lower power, cooling, and management costs

Server performance

Currently, the processor, memory, and I/O subsystem are well balanced and virtually not considered as performance bottlenecks in the majority of systems. The major source of performance issues is related to the storage I/O activity because of the speed of traditional HDD-based storage systems that still does not match the processing capabilities of the servers. As an example, we describe the typical behavior of the OLTP database system.

For database systems, a delicate balance exists between CPU processing power and the I/O throughput needed from the disk. Other factors are involved, such as memory. However, when you add more CPU processing power, you limit I/O because the processor waits for I/O throughput from the disk to proceed to the next instruction. As you add more to your I/O system, the system becomes CPU starved, as depicted in Figure 1.

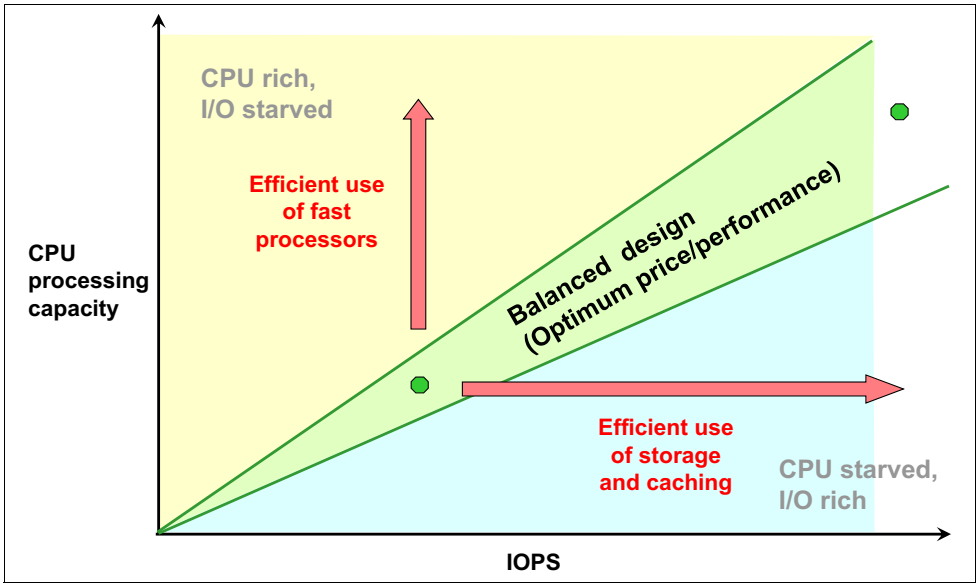

Figure 1 Correlation between CPU processing capacity and I/O throughput

Even a well-tuned database can still experience a substantial I/O wait time. During the time slice depicted in Figure 2, the processor waits at least half of the time for the I/O operations to be processed[1]. If the I/O response time decreases, minimizing processor wait time, the application can process many more operations.

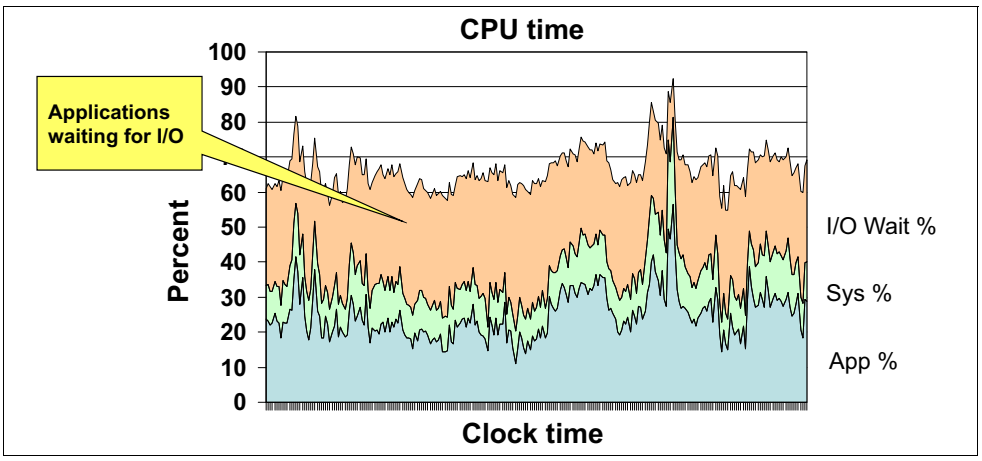

Figure 2 Percent of CPU time spent in I/O Wait compared with Sys % and App % times

The gap between CPU processing power and I/O rates rapidly increased over the last 10 years. This gap is a problem for fast processors that must wait for data to be moved off disks and into memory, and the gap continues to widen.

[1] Here, the I/O Wait % is the time spent waiting for data from the storage system. App % is the time spent running user instructions in a program. System % is time spent managing database locks, shared memory, context switches in memory, and other elements, in support of user programs. See *IBM Information on Demand 2010* by Mike Barton and David Lebutsch, May 2010.

To increase HDD storage access speed, different caching technologies are implemented. Despite the size of the stored dataset itself, only a portion of its data is actively used during normal operations at certain time intervals. The data caching algorithms ensure that the most demanding data (most frequently used) is always kept as close to the application as possible to provide the fastest response time to it.

Caching exists at different levels. Storage controllers use fast dynamic random access memory (DRAM) cache to keep the most frequently used data from disks. However, the cache size in normally limited to several GBs. Operating systems and certain applications keep their own disk cache in the fast system memory, but the cost per GB of RAM storage is very high.

With the introduction of the SSDs, there is an opportunity to dramatically increase the performance of the disk-based storage to match the capabilities of other server subsystems, while keeping costs optimized because the SSDs have lower cost per GB ratio compared to DRAM memory, and lower latency compared to traditional hard drives. This scenario is illustrated in Figure 3.

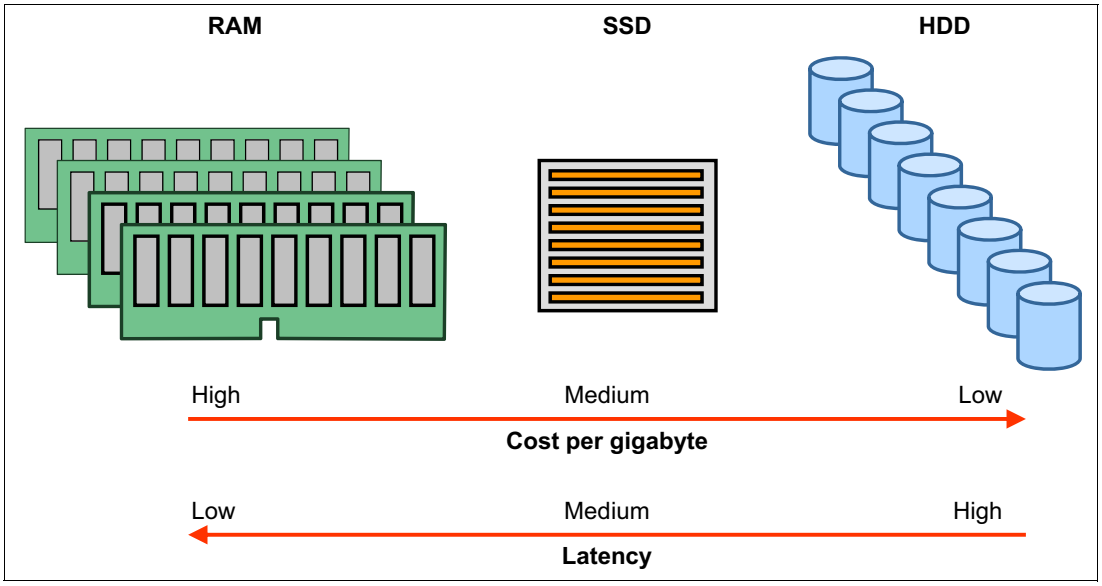

Figure 3 Cost per gigabyte and latency for RAM, SSD, and HDD

SSDs can help fill the gap between processing power and storage I/O. They can help ensure that critical data is moved to the processor or memory much more quickly. SSDs reduce I/O wait time by initiating and completing data operations much more quickly than spinning hard disks. For the highest level of performance, the goal is to keep the processor busy by reducing wait time and spending more time in running operations.

An SSD stores data by using flash memory, which is known as solid-state *Negated AND*, or *NOT AND* (NAND) technology, instead of spinning disks within HDDs. NAND-based memory retains its memory even when the power is cut off. Additionally, NAND-based memory provides much faster access time, latency, and throughput. SSDs use the same interfaces that HDDs use. Therefore, you can easily replace HDDs with SSDs.

In general, there are two key types of storage applications based on workload they generate:

- *I/O-intensive* applications require the storage system to process as many hosts' read and write requests (or I/O requests) per second as possible given the average I/O request size used by this application, typically 8 - 16 KB. This behavior is most common for OLTP databases.

- *Throughput-intensive* applications require the storage system to transfer to or from hosts as many gigabytes of information per second as possible, and they typically use an I/O request size of 64 - 128 KB. These characteristics are commonly inherent to file servers, multimedia streaming, and backup.

Therefore, there are two key performance metrics to evaluate storage system performance: *input/output requests per second (IOPS)* and *throughput (measured in MBps or GBps)*, depending on application workload.

Another important factor to take into account is the *response time (or latency)*, or how much time does the application spend waiting for the response from the storage system after submitting a particular I/O request. In other words, response time is the amount of time required by the storage system to complete an I/O request. Response time has a direct impact on the productivity of users who work with the application, such as how long it takes to get the requested information, and on the application itself. For example, a slow response to the database write requests might cause multiple record locks and further performance degradation of the application.

Key factors affecting the response time of the storage system itself include, how quickly required data can be located on the media (seek time), and how quickly they can be read from or written to the media. That is, response time also depends on the size of the I/O request (reading or writing more data normally takes more time).

In addition, the majority of applications generate many storage I/O requests at the same time, and these requests might spend some time in the queue if they cannot be immediately handled by the storage system. The number of I/O requests that can be concurrently sent to the storage system for the execution is called *queue depth*. This represents the service queue; that is, the queue of requests that is currently being processed by the storage system. If the number of outgoing I/O requests outreaches the parallel processing capabilities of the storage system (I/O queue depth), the requests are put into the wait queue, and then moved to the service queue when a spot becomes available. This also affects the overall response time.

From the traditional spinning HDD perspective, improvement of its latency is limited by mechanical design. Despite the increase in rotational speed of the disk plate and density of stored data, the response time of the HDD is still several milliseconds, which effectively limit its maximum IOPS (for example, single 2.5-inch 15 K rpm SAS HDD is capable of ~300 IOPS).

With the SSD-based eXFlash, the latency is measured in dozens of microseconds (or almost 100 times lower than for the hard drives), which in turn leads to more than 200,000 IOPS in a RAID-5 configuration for a single eXFlash unit. Higher IOPS capabilities also mean higher queue depth and therefore, better response time for almost all types of storage I/O-intensive applications.

In other words, if the application is multi-user, heavy loaded, and has access storage with random I/O requests of a small size, this application is a very good candidate to consider putting its entire data set or part of it on eXFlash. Vice versa, if an application transfers large amounts of data, such as backups or archiving, eXFlash might not provide any advantage because the limiting factor will be the bandwidth of SSD interface.

The knowledge of how the application accesses data, such as read-intensive or write-intensive, and random data access or sequential data access, helps to implement the most cost-efficient storage that meets required service level agreement (SLA) parameters. Table 1 summarizes typical application workload patterns that are suitable for eXFlash-based deployments in multi-user environments, depending on application type.

Table 1 Typical application workload patterns

Workload type → ↓ Application type	**Read intensive**	**Write intensive**	**I/O intensive**	**Throughput intensive**	**Random access**	**Sequential access**	**Good for eXFlash**
OLTP database	Yes	Yes	Yes		Yes		Yes
Data warehouse	Yes			Yes	Yes		Yes
E-mail server	Yes	Yes	Yes		Yes		Yes
Medical imaging	Yes			Yes	Yes		Yes
Video on demand	Yes			Yes	Yes		Yes
Web/Internet	Yes		Yes		Yes		Yes
Web 2.0	Yes	Yes	Yes		Yes		Yes

There are two types of SSDs that can be used with eXFlash: *Enterprise SSDs* and *Enterprise Value SSDs*. Although both Enterprise SSDs and Enterprise Value SSDs typically have similar read IOPS characteristics, the key difference between them is their write IOPS performance and their endurance (or life expectancy).

SSDs have a huge, but finite number of program/erase (P/E) cycles, which affect how long they can perform write operations and thus their life expectancy. Enterprise SSDs have significantly better endurance but a higher cost/IOPS ratio compared to Enterprise Value SSDs. SSD write endurance is typically measured by the number of program/erase cycles that the drive can incur over its lifetime, which is listed as TBW (Total Bytes Written) in the device specification.

Because of such behavior by Enterprise Value solid-state drives, careful planning must be done to use them only in read-intensive environments to ensure that the TBW of the drive will not be exceeded before the required life expectancy.

In other words, for an optimal cost/IOPS ratio, you should use Enterprise SSDs for write-intensive workloads, and Enterprise Value SSDs for read-intensive workloads, as shown in Table 2.

Table 2 Application workload by SSD type

SSD type → ↓ Application type	**Enterprise**	**Enterprise Value**	**Good for eXFlash**
OLTP database	Yes		Yes
Data warehouse		Yes	Yes
E-mail server	Yes		Yes
Medical imaging		Yes	Yes
Video on demand		Yes	Yes
Web/Internet		Yes	Yes
Web 2.0	Yes		Yes

Typical eXFlash usage scenarios include:

- High-speed read cache in a local or SAN-based storage environment
- Temporary local storage space for mid-tier applications and databases
- Main (Tier 0) local data storage in a single server environment or in a distributed scale-out environment with local-only storage or mixed local and SAN-based storage

eXFlash technology

eXFlash technology is a server-based high performance internal storage solution, which is based on SSDs and performance-optimized disk controllers (both Redundant Array of Independent Disks (RAID) and non-RAID).

A single eXFlash unit accommodates up to eight hot-swap SSDs, and can be connected to up to two performance-optimized controllers. eXFlash is supported on System x3690 X5, x3850 X5, x3950 X5, x3750 M4, and x3650 M4 servers.

Figure 4 shows an eXFlash unit, with the status lights assembly on the left side.

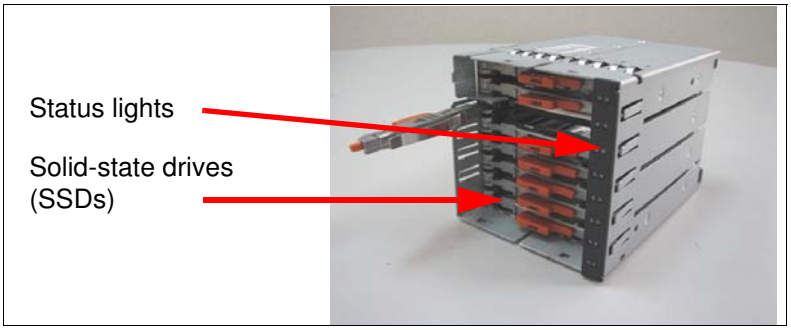

Figure 4 eXFlash unit

Each eXFlash unit can accommodate up to eight 1.8-inch hot-swap front-accessible SSDs, and it occupies four 2.5-inch SAS hard disk drive bays. You can install the following number of eXFlash units:

- The x3850 X5 can have up to sixteen 1.8-inch SSDs with up to two eXFlash units (up to eight SSDs per eXFlash unit).
- The x3690 X5 can have up to twenty-four 1.8-inch SSDs with up to three eXFlash units (up to eight SSDs per eXFlash unit).
- The x3750 M4 and x3650 M4 can have up to thirty-two 1.8-inch SSDs with up to four eXFlash units (up to eight SSDs per eXFlash unit).

In theory, the OLTP I/O performance of the single eXFlash unit combined with the S3700 MLC Enterprise SSDs configured in a RAID-5 array, is equivalent to the storage system consisting of more than 1000 traditional spinning HDDs. Besides HDDs themselves, building such a massive I/O-intensive high-performance storage system requires external deployment with many additional infrastructure components including host bus adapters (HBAs), switches, storage controllers, disk expansion enclosures, and cables. Consequently, this leads to more capital expenses, floor space, electrical power requirements, and operation and support costs. Because eXFlash is based on internal server storage, it does not require all those components discussed above, and helps to eliminate additional expenses and environmental requirements.

In summary, the eXFlash solution with the S3700 or S3500 SSDs can provide the following benefits:

- Significantly lower implementation costs of high performance I/O-intensive storage systems with the best cost per IOPS ratio
- Significantly higher performance and better response time of storage-intensive applications with up to ten times less response time
- Significant savings in power and cooling with high performance per watt ratio
- Significant savings in floor space with extreme performance per unit of the rack space ratio
- Simplified management and maintenance with internal server-based configurations (no external power and information infrastructure needed)

eXFlash is optimized for a heavy mix of random read and write operations, such as transaction processing, data mining, business intelligence, and decision support, and other random I/O-intensive applications. In addition to its superior performance, eXFlash offers superior uptime with three times the reliability of mechanical disk drives. SSDs have no moving parts to fail. They use Enterprise Wear-Leveling to extend their use even longer.

In environments where RAID protection is required, that is, eXFlash is used as a master data storage, use a RAID controller with the Performance Accelerator key enabled to ensure the peak IOPS can be reached.

The main advantage of M5014, M5015, M5016, and M5110 with Performance Key controllers for SSDs is a Cut Through I/O (CTIO) feature enabled. CTIO optimizes highly random read and write I/O operations for small data blocks to support the high IOPS capabilities of SSD drives and to ensure the fastest response time to the application. For example, enabling CTIO on a RAID controller with SSDs allows you to achieve up to two times more IOPS compared to the controller with the CTIO feature disabled.

Note: A single eXFlash unit requires a dedicated controller (or two controllers). When used with eXFlash, these controllers cannot be connected to the HDD backplanes.

In a non-RAID environment where eXFlash can be used as a high-speed read cache, use the 6 Gb Performance Optimized HBA to ensure that maximum random I/O read performance is achieved. Only one 6 Gb SSD HBA is supported per single SSD backplane.

It is possible to mix RAID and non-RAID environments; however, the maximum number of disk controllers that can be used with all SSD backplanes in a single system is *four*.

eXFlash requires the following components:

- eXFlash hot-swap SAS SSD backplane
- 1.8-inch solid-state drives (SSDs)
- Supported disk controllers

S3700 SATA MLC Enterprise SSDs for System x

The S3700 MLC Enterprise solid-state drives represent next generation enterprise-grade SSD technology that is based on a 25 nm High Endurance Technology (HET) MLC NAND flash. These drives provide robust endurance: one drive can be fully rewritten up to ten times per day. This makes these drives particularly suitable for heavy-loaded write-intensive application environments such as databases, corporate email, Web 2.0, and others.

The S3700 MLC SATA Enterprise SSD is shown in Figure 5.

Figure 5 S3700 SATA MLC Enterprise SSD (2.5 inch shown)

S3700 MLC Enterprise SSDs are available in 1.8-inch and 2.5-inch form factors. However, eXFlash supports 1.8-inch SSDs only; therefore, we discuss only 1.8-inch SSDs in this publication.

S3700 MLC Enterprise SSDs have the following key features:

- Fast and consistent performance

  Deliver data at a breakneck pace, with consistently low latencies, and tight IOPS distribution, boosting CPU utilization.

- Stress-free protection

  End-to-end data protection provides multiple secure checkpoints to catch any data errors, wherever they occur, including full data path protection, parity, cyclic redundancy check (CRC), memory error correction code (ECC), and logical block address (LBA) tag validation.

- High-endurance technology

  Meet your most demanding needs with marathon-like write endurance of up to 7.3 PB TBW for 400 GB SSD.

Table 3 summarizes specifications of S3700 SATA 1.8-inch MLC Enterprise SSDs.

Table 3 S3700 SATA 1.8-inch MLC Enterprise SSD specifications

Specification	**200 GB**	**400 GB**
Interface	6 Gbps SATA	6 Gbps SATA
Hot-swap drive	Yes	Yes
Form factor	1.8-inch	1.8-inch
Capacity	200 GB	400 GB
Endurance, TBW	3.65 PB	7.3 PB
IOPS read[a]	75,000	75,000
IOPS write[a]	29,000	36,000
Sequential read rate[b]	500 MBps	500 MBps
Sequential write rate[b]	365 MBps	460 MBps

Specification	200 GB	400 GB
Read latency	50 µs	50 µs
Write latency	65 µs	65 µs
Typical power	6 W	6 W

a. 4 KB block transfers
b. 128 KB block transfers

For more information, refer to the *S3700 MLC Enterprise SSDs Product Guide*, at the following website:

http://lenovopress.com/tips1014

S3500 SATA MLC Enterprise Value SSDs for System x

The S3500 MLC Enterprise Value solid-state drives represent leading edge 20 nm MLC NAND flash technology. Unlike client solid-state drives, S3500 SATA MLC Enterprise Value SSDs are designed to operate 24 hours per day, 7 days per week (24x7). They are equipped with a robust suite of enterprise features, including specific, measurable, achievable, relevant, and timely (SMART) attributes; hot-plug support; high reliability; enhanced ruggedness; thermal throttling; and enhanced power loss protection. They also use full end-to-end data path protection to protect the integrity of the data that is transferred to and from the NAND flash memory or stored in the memory.

S3500 SATA MLC Enterprise Value SSDs provide outstanding IOPS/watt and cost/IOPS for read-intensive enterprise applications such as web serving, content delivery, streaming video, and big data.

The S3500 MLC SATA Enterprise Value SSD is shown in Figure 6.

Figure 6 S3500 SATA MLC Enterprise Value SSD (2.5 inch shown)

S3500 MLC Enterprise Value SSDs are available in 1.8-inch and 2.5-inch form factors; however, eXFlash supports 1.8-inch SSDs only. Therefore, we discuss only 1.8-inch SSDs in this publication.

S3500 MLC Enterprise Value SSDs have the following key features:

- Fast and consistent performance

  Deliver data at a breakneck pace, with consistently low latencies, and tight IOPS distribution, boosting CPU utilization.

- Stress-free protection

  End-to-end data protection provides multiple secure checkpoints to catch any data errors, wherever they occur, including full data path protection, parity, CRC, memory ECC, and LBA tag validation.

Table 4 summarizes the specifications of S3500 SATA 1.8-inch MLC Enterprise Value SSDs.

Table 4 S3500 SATA 1.8-inch MLC Enterprise Value SSD specifications

Specification	80 GB	240 GB	400 GB
Interface	6 Gbps SATA	6 Gbps SATA	6 Gbps SATA
Hot-swap drive	Yes	Yes	Yes
Form factor	1.8-inch	1.8-inch	1.8-inch
Capacity	80 GB	240 GB	400 GB
Endurance, TBW	45 TB	140 TB	225 TB
IOPS read[a]	70,000	75,000	75,000
IOPS write[a]	7,000	7,500	11,000
Sequential read rate[b]	340 MBps	500 MBps	500 MBps
Sequential write rate[b]	100 MBps	260 MBps	380 MBps
Read latency	50 µs	50 µs	50 µs
Write latency	65 µs	65 µs	65 µs
Typical power	5 W	5 W	5 W

a. 4 KB block transfers
b. 128 KB block transfers

For more information, refer to the *S3500 MLC Enterprise Value SSDs Product Guide*, found at the following website:

http://lenovopress.com/tips1067

OLTP database deployment scenario

For illustration purposes, we describe how the eXFlash unit combined with S3700 MLC Enterprise SSDs could potentially perform in a typical online transaction processing (OLTP) database solution compared to the traditional HDD-based approach.

OLTP is a multi-user, memory-, CPU-, and storage I/O-intensive, random workload. It is characterized by many small read and write storage I/O requests (typically four or eight kilobytes and 70/30 read/write ratio) that are generated by transactions originated by multiple users. The transactions are relatively simple; however, every transaction can generate dozens of physical storage I/O requests depending on transaction type, application architecture, and business model used.

The key performance indicator of transactional systems is the response time because the client expects to get the requested product information or to place an order very quickly. If the expectations are not met, the chance that the client will go to the competitor is very high. Because of that, storage IOPS performance is considered as a very important factor to ensure that the response time goals are met, and to keep other system resources (CPU, memory) under good utilization and not waiting for the data.

Assume that we need 150,000 IOPS of storage I/O performance for our OLTP environment, and we use System x3850 X5™ as our database server. We evaluate two approaches: traditional with spinning HDDs, and eXFlash with S3700 MLC Enterprise SSDs.

For the purpose of our calculation, we assume the following parameters:

- OLTP workload (random 70/30 read/write operations mix), 4 KB block transfers, RAID-5 array
- One eXFlash unit with eight S3700 MLC Enterprise SSDs is theoretically projected to deliver 235,000 IOPS under defined workload
- One traditional HDD is estimated to deliver 175 IOPS under defined workload

Figure 7 illustrates this scenario using two approaches: Achieving required IOPS with traditional HDDs, and with eXFlash.

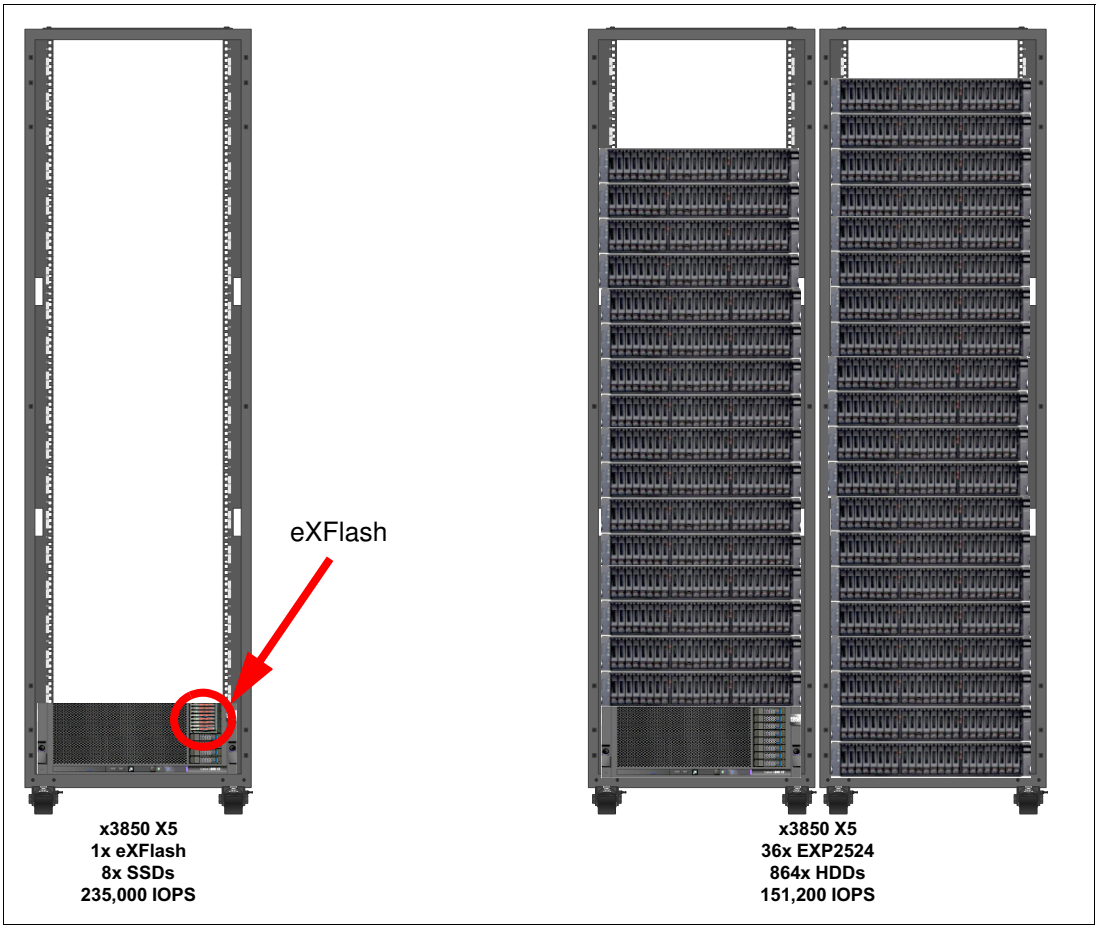

Figure 7 eXFlash versus traditional HDDs: OLTP databases

Using the traditional HDD-based approach, we need to deploy one x3850 X5 connected to 36 IBM System Storage EXP2524 with 864 hard drives (24 drives per EXP2524 enclosure), by four ServeRAID™ M5120 SAS/SATA controllers.

With eXFlash and S3700 SSDs, which are capable of more than 200,000 IOPS for OLTP workloads, there is no need for external disk storage at all, because all I/O performance requirements are met internally. Although both configurations can handle 150,000 IOPS, an eXFlash-based configuration has a higher theoretical IOPS capacity of 235,000 IOPS, plus the ability to add a second eXFlash unit for a total of 470,000 IOPS. However, an HDD-based configuration has an absolute maximum of 151,200 IOPS for the given number of drives (this is the maximum configuration and cannot be scaled further). In addition, the response time of an eXFlash-based configuration is significantly better.

As you can see, eXFlash is able to significantly reduce power and cooling requirements, occupied rack space, management, and operational costs while providing better reliability and the same or better performance levels.

It is also important to mention that the increased system utilization potentially leads to fewer systems that need to be deployed, therefore, providing additional management and software licensing cost savings.

Table 5 summarizes characteristics of each scenario and highlights eXFlash advantages.

Table 5 Traditional HDDs versus eXFlash: OLTP workload scenario

Characteristic	Traditional HDD	eXFlash	eXFlash advantage
Number of servers	1	1	Fewer components to acquire and maintain, higher reliability, server-based management
Number of drives	864	8	
Drive type	73 GB 15 K 2.5-in. SAS HDD	S3700 400 GB 1.8-in. SATA SSD	
Location	External	Internal	
External enclosures	36	0	
RAID controllers	4	1	
External cables	36	0	
Maximum IOPS capacity	151,200	235,000	Transparent performance scaling at no additional cost, significantly better cost per IOPS ratio
Used IOPS capacity	150,000	150,000	
Raw storage capacity	63 TB	3.2 TB	Significantly better storage space utilization, no wasted storage space, comparable cost per utilized GB ratio
Effective RAID capacity	58.7 TB	2.8 TB	
Used capacity	2 TB	2 TB	
Power consumption (unit of measurement)	kW	W	Significant savings in power and cooling costs, no external power infrastructure required
Rack U space	76U (2x racks)	4U	No additional rack space required, smaller footprint

To ensure that the high availability requirements are met in case of node failure, several techniques can be utilized depending on database vendor. Such techniques include:

- Log shipping
- Replication
- Database mirroring

The partitioning feature of many databases, for example, IBM DB2, can help to split the workload between several nodes therefore increasing overall performance, availability, and capacity.

If, for some reason, the entire database cannot be placed onto eXFlash, consider placing at least some of the data there. The areas to look at include:

- Log files
- Temporary table space
- Frequently-accessed tables
- Table partitions
- Indexes

Some databases like Oracle support extension of their own data buffers to the SSDs, which provides a significant cost-efficient performance increase.

There are more complex configurations, where tiered storage is implemented with both internal and external storage, and the data is moved between storage tiers based on defined policies. The data movement can be implemented manually using the available database tools (like DB2 Reorg utility, in DB2), or automatically, by using the respective storage management software, for example, IBM GPFS.

Video-on-demand deployment scenario

In a multi-user environment, video-on-demand is a read-intensive and throughput-intensive random workload, where every user receives their own data stream watching different content or even the same content with some delay (for example, a recently published new movie). This requires faster response time to ensure a better user experience and smoother video playback. In general, video streaming applications use larger I/O blocks (64 KB or more) to interact with the storage system.

Consider the following scenario. A provider of on-demand video content needs to support up to 20,000 active users, and they watch 2,000 videos simultaneously (that is, one video is simultaneously watched by 10 active users, on average). The provider uses High Definition (HD) video that requires about 8 Mbps or 1 MBps per stream, and the average movie size is 5 GB.

Assume that a single System x3690 X5 server used in this scenario is capable of handling 4,000 simultaneous video streams. Therefore, we need five servers to support our projected workload. The total throughput for 4,000 concurrent streams is 4 GBps per server, and each server stores 400 videos, which requires 2 TB of storage space.

Based on Table 2 on page 7, we selected S3500 Enterprise Value SSDs for our workload.

For our calculation, we assume the following parameters:

- One eXFlash unit with eight S3500 MLC Enterprise SSDs is theoretically projected to deliver up to 4 GBps throughput under random read workload
- One EXP2524 with 24 traditional HDDs is estimated to deliver up to 1 GBps throughput under random read workload

Figure 8 illustrates this scenario using two approaches: Achieving required throughput with traditional HDDs, and with eXFlash.

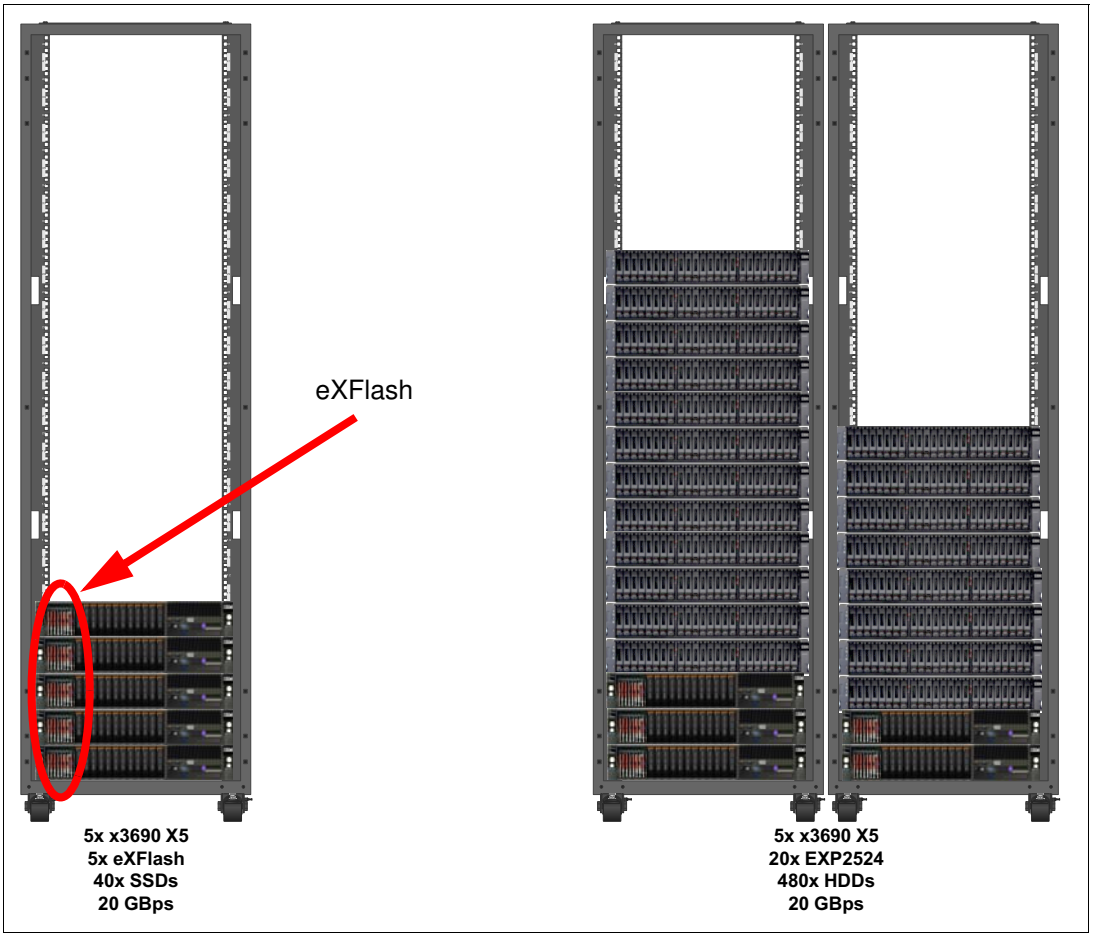

Figure 8 eXFlash versus traditional HDDs: video-on-demand

With the traditional HDD scenario, to achieve such a throughput of 4 GBps per server with random I/O reads, we need about 96x 73 GB 15 K rpm SAS HDDs in four EXP2524 external storage expansion units per server for a total of 480 HDDs and 20 EXP2524 units. With the eXFlash scenario, we need one eXFlash unit with eight S3500 Enterprise Value SSDs per server to meet the capacity and throughput requirements.

Table 6 summarizes characteristics of each scenario and highlights eXFlash advantages.

Table 6 Traditional HDDs versus eXFlash: video-on-demand workload scenario

Characteristic	Traditional HDD	eXFlash	eXFlash advantage
Number OS servers	5	5	Fewer components to acquire and maintain, higher reliability, server-based management
Number of drives	480	40	
Drive type	73 GB 15 K 2.5-in. SAS HDD	S3500 400 GB 1.8-in. SATA SSD	
Location	External	Internal	
External enclosures	20	0	
RAID controllers	10	5	
External cables	20	0	
Maximum throughput	20 GBps	20 GBps	Same throughput with fewer components and better response time
Used throughput	20 GBps	20 GBps	
Raw storage capacity	35 TB	16 TB	Significantly better storage space utilization, no wasted storage space, comparable cost per utilized GB ratio
Effective RAID capacity	31.2 TB	14 TB	
Used capacity	10 TB	10 TB	
Power consumption (unit of measurement)	kW	W	Significant savings in power and cooling costs, no external power infrastructure required
Rack U space	50U (2x racks)	10U	No additional rack space required, smaller footprint

Conclusion

The growth in CPU processing power far exceeds the growth in disk I/O. For this reason, disk I/O is the culprit in major bottlenecks in many high-performance applications. The eXFlash with S3700 or S3500 solid-state drives makes it possible for you to eliminate I/O bottlenecks in storage-intensive enterprise workloads. Compared with HDDs, the eXFlash eliminates I/O bottlenecks at a much lower cost while maintaining the same level of performance, thus making it a more cost-effective workload-optimized system.

The eXFlash solution offers a dramatic increase in I/O throughput for storage-intensive enterprise workloads. For scalable servers, such as the System x3690 X5 or x3850 X5, adding eXFlash SSDs is a cost-effective way to add needed extra I/O throughput, rather than using multiple hard disks in the external storage systems.

Our sample deployment scenarios highlighted that the eXFlash with S3700 or S3500 solid-state drives can be a feasible cost-optimized alternative to the traditional HDD-based storage architectures. These drives, combined with the eXFlash, can help achieve higher reliability and availability of the business critical data with faster access to it, lower acquisition costs with a fewer number of devices and components, shorten the return on investment (ROI) time frame, and decrease overall total cost of ownership (TCO) with the efficient utilization of server resources and lower power, cooling, and management costs.

Authors

This paper was produced by the following team of specialists:

Ilya Krutov is a Project Leader at Lenovo Press. He manages and produces pre-sale and post-sale technical publications on various IT topics, including x86 rack and blade servers, server operating systems, virtualization and cloud, networking, storage, and systems management. Ilya has more than 15 years of experience in the IT industry, backed by professional certifications from Cisco Systems, IBM, and Microsoft. During his career, Ilya has held a variety of technical and leadership positions in education, consulting, services, technical sales, marketing, channel business, and programming. He has written more than 200 books, papers, and other technical documents. Ilya has a Specialist's degree with honors in Computer Engineering from the Moscow State Engineering and Physics Institute (Technical University).

Thanks to the following people for their contributions to this project:

Randall Lundin
Lenovo

Brad Buland
Intel

Thanks to the authors of *The Benefits of Optimizing OLTP Databases Using IBM eXFlash Solid-State Drives*, REDP-4849: Tim Bohn, Dave Pierson, Brian Haan.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/redp4948

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

eXFlash™	Lenovo(logo)®	X5™
Lenovo®	ServeRAID™	

The following terms are trademarks of other companies:

Intel, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.