Ilya Krutov

Benefits of Lenovo eXFlash Memory-Channel Storage in Enterprise Solutions

The increasing demand for cloud computing and business analytical workloads by enterprises to meet their business needs drives innovation to find new ways to build their informational systems. Clients are looking for cost-optimized fit-for-purpose IT solutions that manage large amounts of data, easily scale performance, and provide reliable real-time access to the actionable information. 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 many performance-demanding applications.

The Lenovo® eXFlash™ memory-channel storage is an innovative high performance solid-state storage device that closes the disk I/O performance gap by connecting the flash memory module directly to a DDR3 memory bus using a standard DIMM form factor.

The eXFlash DDR3 Storage DIMMs offer ultralow latency, highly scalable storage technology that helps decrease overall total cost of ownership (TCO) with the efficient use of server resources, faster response time, and lower software, power, cooling, and management costs.

This paper describes server performance imbalance that can be found in typical application environments. It also describes how to address the issue with the eXFlash DIMMs to provide required levels of scalability, performance, and availability for the storage-intensive applications.

The following topics are covered:

Executive summary

Currently, the processor and memory subsystems are well-balanced and virtually not considered as performance bottlenecks in most server 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 (both server internal and external shared storage) that still does not match the processing capabilities of the servers. This disparity can lead to a situation where a powerful processor sits idle waiting for the storage I/O requests to complete. This situation wastes the processor's time, which negatively affects user productivity, extends the time frame of return on investments (ROI), and increases overall total cost of ownership (TCO).

With the virtualization trends in data centers, servers demand higher I/O throughput to match the capabilities of multi-core processors and increased amounts of memory, allowing the higher number of virtual machines (VMs) to be hosted on a single physical system. Higher I/O throughput, including storage I/O, can help achieve better server utilization and a higher number of VMs per server.

With transactional databases, higher storage I/O throughput can help improve user and business productivity by lowering overall response time. Data warehouses and business analytics are other examples of the workload that requires higher storage I/O throughput to allow faster data processing, making strategic business decisions in a timely manner.

Lenovo eXFlash memory-channel storage, an ultralow latency, highly scalable storage technology, can help significantly decrease storage I/O response time to match the processing power of the server CPUs, decreasing overall TCO with efficient use of server resources, faster response time, and lower software, power, cooling, and management costs.

Lenovo eXFlash memory-channel storage can help achieve these goals:

- The highest IOPS density per GB and the lowest write latency among Lenovo flash storage offerings for the most demanding IOPS-intensive and latency-sensitive applications.
- More than 10 times faster storage write-access compared to traditional solid-state drives (SSDs).
- Almost linear storage I/O performance scalability for both read- and write-intensive enterprise applications.
- Ability to virtualize data-intensive enterprise applications that could not be virtualized because of storage I/O constraints.
- Higher reliability and availability of services because of the fewer number of components that are used to build the solution.
- Lower acquisition costs because of fewer number of systems and components.
- Shorten ROI time frame and decrease overall TCO with the efficient utilization of server resources and lower software, power, cooling, and management costs.

Introduction

Storage is no longer an afterthought. Companies are searching for more ways to efficiently manage expanding volumes of data and to make that data accessible throughout the enterprise. In addition, strategic workloads, such as data analytics, which support business decisions, rely on efficient data mining, including data access speed and processing time. Rapid adoption of virtualization and cloud technologies in data centers also drives the need for a high-speed storage I/O throughput.

With current estimates of the amount of data to be managed and made available, increasing significantly each year, and tight budgets and spending cuts, there is a need to implement efficient storage infrastructure that can help lower TCO while providing required levels of performance, scalability, flexibility, availability, and improved data access.

However, from the system performance point of view, the processor and memory subsystems are typically 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. As an example, consider the typical behavior of the online transaction processing (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.

This balance is illustrated in Figure 1.

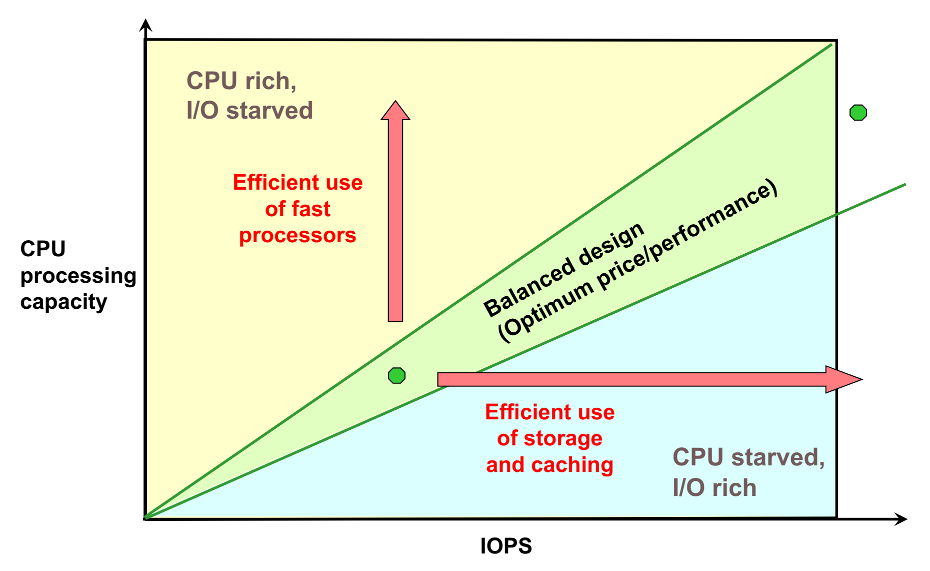


Figure 1 Correlation between CPU processing capacity and I/O throughput

Even a well-tuned database can still experience a substantial I/O wait time. During the time slice depicted in Figure 2 on page 4, the processor waits at least half of the time for the I/O operations to be processed. If the I/O response time decreases, minimizing processor wait time, the application can process many more operations.

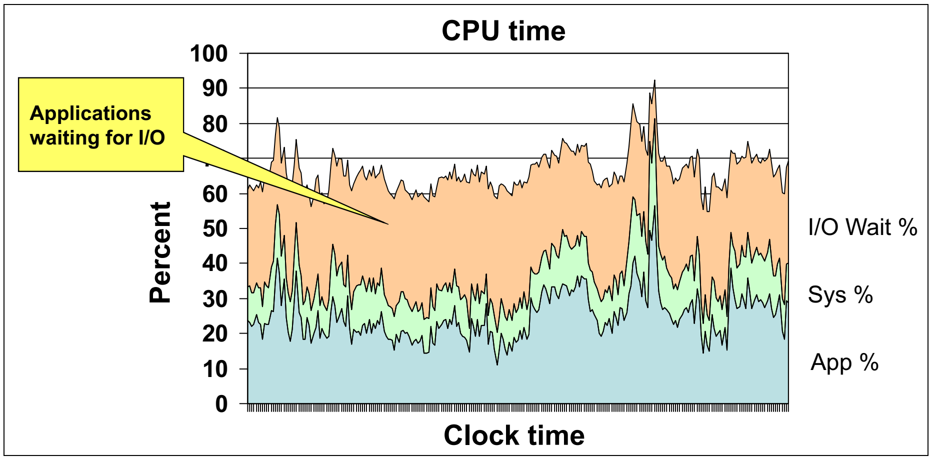


Figure 2 Percent of CPU time spent in I/O Wait compared with Sys % and App % times[1]

The gap between CPU processing power and I/O rates has 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. The performance gap drives the need to find the new ways of storage implementation that can help shorten or eliminate the performance imbalance. There are two primary approaches to address the issue:

- Placing the entire data set onto a higher speed main (permanent) storage
- Using the higher speed storage as a temporary buffer (cache) to store a subset of the entire data set

Figure 3 compares data access speed (latency) and cost per gigabyte for various storage types that are currently available.

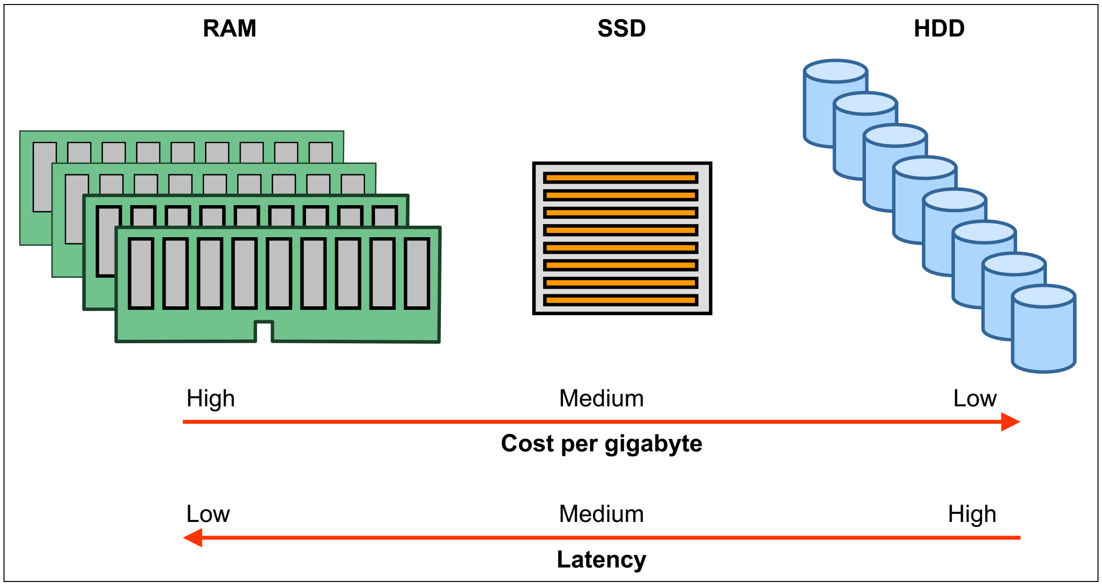


Figure 3 Cost per gigabyte and latency for RAM, solid-state storage, and HDD

[1] Here, the I/O Wait % is the time spent waiting for data from the storage system. The App % is the time spent running user instructions in a program. The 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.

Keeping data in fast RAM ensures the fastest access speed. RAM as a main storage is typically used with in-memory databases; RAM as a disk cache is commonly used by the operating systems and applications. However, the cost per GB of RAM storage is high.

With the introduction of the solid-state storage, such as Lenovo eXFlash memory-channel storage, solid-state drives (SSDs), and PCIe SSD adapters, 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 flash storage has lower cost per gigabyte ratio compared to DRAM memory, and significantly lower latency compared to traditional hard disk drives (HDD).

Solid-state storage can help fill the gap between processing power and storage I/O. It can help ensure that critical data is moved to the processor or memory much more quickly. Flash storage reduces 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.

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

- *IOPS-intensive* applications require the storage system to process as many read and write requests (or I/O requests) per second of hosts as possible, given the average I/O request size used by this application, which is typically 4 - 8 KB. This behavior is most common for OLTP databases.
- *Bandwidth-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 128 - 512 KB or more. These characteristics are commonly inherent to file servers, media streaming, and backup.

Therefore, two key performance metrics can evaluate storage system performance, depending on application workload:

- Input/output requests per second (IOPS)
- Throughput (measured in MBps or GBps)

Another important factor to consider is the *response time* (or *latency*), which is how much time the application spends waiting for the response from the storage system after submitting a particular I/O request. Another way to say it is that 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 HDD-based storage system 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 subsystem. 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 15,000 rpm SAS HDD is capable of ~300 IOPS using 4 KB blocks).

With the SSD-based storage, the latency is typically measured in dozens of microseconds (or almost 100 times lower than for the HDDs), which in turn leads to significantly higher IOPS per solid-state device (typically, ~50,000 IOPS or more). Higher IOPS capabilities also mean higher queue depth and therefore, better response time for almost all types of storage I/O-intensive applications.

If the application is multi-user, heavy loaded, and has access storage with random I/O requests, this application is a good candidate to consider for SSD device placement.

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 the placement onto flash storage in multi-user environments, depending on application type.

Table 1 Typical application workload patterns

Workload type → ↓ Application type	**Read intensive**	**Write intensive**	**IOPS intensive**	**Bandwidth intensive**	**Rand. access**	**Seq. access**	**Latency sensitive**	**Good for SSD**
OLTP database	Yes	Yes	Yes		Yes		Yes	Yes
Data warehouse	Yes			Yes	Yes		Yes	Yes
File server	Yes			Yes	Yes			
Email server	Yes	Yes	Yes		Yes		Yes	Yes
Medical imaging	Yes			Yes	Yes		Yes	Yes
Video on demand	Yes			Yes	Yes		Yes	Yes
Web/Internet	Yes		Yes		Yes		Yes	Yes
Web 2.0	Yes	Yes	Yes		Yes		Yes	Yes
Archives/backup		Yes		Yes		Yes		

Lenovo eXFlash memory-channel storage is an industry-first solid-state storage technology that uses a standard DIMM form factor to significantly increase server and storage performance and efficiency, helping eliminate storage I/O bottlenecks for most of workloads that are listed in Table 1.

Benefits

The primary source of performance issues in server-based systems tends to be related to storage I/O activity because the speed of traditional storage systems still does not match the processing capabilities of the servers. Therefore, increasing the storage I/O throughput can help close or minimize this performance gap and ensure that performance optimization is realized for other system resources (processor and memory).

Impact of slow storage I/O

In general, you might experience the following challenges in storage I/O-constrained environments:

- Failure to meet user expectations and service levels because of slow application response time
- Decreased user and business productivity
- Application and data availability concerns (slow batch processing, long backup windows, and hardware failure rates)
- Increased storage performance requirements
- Scalability constraints because of data center space, power, and cooling limits
- Increased TCO:
  - Rising data center power and cooling costs
  - Increasing software licensing fees
  - Rising server, network, and storage infrastructure management and support costs
  - Longer lead time to ROI because of inefficient utilization of the existing resources

Lenovo eXFlash memory-channel storage benefits

Lenovo eXFlash memory-channel storage can help address the challenges in the following ways:

- Dramatically boosts the performance of existing applications while lowering cost per IOPS ratio
- Increases user productivity with better response times, improving business efficiency
- Increases storage performance while decreasing power, cooling, and space requirements
- Reduces TCO:
  - Reduces energy costs because of lower power and cooling requirements
  - Reduces the number of systems, devices, and components that are required to build the solution by increasing usage of available resources
  - Reduces software license fees because fewer systems or processors are required
  - Reduces management and support costs because of fewer components to deploy
  - Faster ROI because of better resource usage

In summary, the use of eXFlash DIMMs can help achieve the following benefits:

- Higher performance:
  - Higher IOPS
  - Higher throughput
  - Lower latency
- Infrastructure simplification:
  - Higher number of virtual machines (VMs) and user density
  - Lower number of physical systems
  - Simplified deployment and management
- Improved TCO:
  - Reduced acquisition costs
  - Reduced software licensing fees
  - Reduced power and cooling costs
  - Reduced support and maintenance costs

In this paper, we describe the following possible use cases where eXFlash DIMMs can provide certain benefits:

- Transactional databases (OLTP)
- Data warehouses (OLAP)
- Virtualized workloads

See "Use cases" on page 14 for further information.

Positioning

The following flash storage offerings can help achieve fast response time for analytical workloads, transactional databases, and virtualized environments:

- Lenovo eXFlash memory-channel storage (eXFlash DIMMs): A perspective innovative flash storage solution that uses DDR3 memory channels to connect flash storage modules.
- eXFlash: Innovative high-density design of the drive cages and the performance-optimized storage controllers with the reliable high-speed solid-state drive technology.
- High IOPS MLC Adapters: Utilize the latest enterprise-level solid-state storage technologies in a standard PCIe form factor and include sophisticated advanced features to optimize flash storage and help deliver consistently high levels of performance, endurance, and reliability.
- 2.5-inch Enterprise SSDs: Designed to be flexible across a wide variety of enterprise workloads in delivering outstanding performance, reliability, and endurance at an affordable cost.
- FlashCache™ Storage Accelerator: All-in-one flash-caching product that leverages the speed, management, capacity and breadth of the Lenovo qualified solid-state storage, integrating them into a high speed server-side caching service that seamlessly accelerates the most important data with little or minimal IT overhead in both physical and virtual servers.

Table 2 compares features of the flash storage devices.

Table 2 Lenovo flash storage devices

Feature	eXFlash DIMMs	eXFlash SSDs	PCIe SSD adapters	2.5-inch SSDs
Form factor	LP DIMM	1.8-inch drive	PCIe adapter	2.5-inch drive
Interface	DDR3 1600 MHz	6 Gbps SATA	PCIe 2.0 x8	6 Gbps SAS or SATA
Capacity	Up to 400 GB	Up to 400 GB	Up to 2.4 TB	Up to 1.6 TB
Maximum random read IOPS	More than 135,000	Up to 75,000	Up to 285,000	Up to 100,000
Write latency	Less than 5 µs	65 µs	15 µs	Less than 100 µs
Hot-swap capabilities	No	Yes	No	Yes
RAID support	No	Yes	Chip-level redundancy	Yes

As a general consideration, 2.5-inch SSDs can be an entry-level solution that is optimized for commodity servers with conventional HDD drive tray, with moderate storage IOPS performance requirements.

Both PCIe SSD adapters and eXFlash SSDs are optimized for the storage I/O-intensive enterprise workloads, where PCIe adapters offer significantly lower write latency and the eXFlash SSDs offer better IOPS density, and leverage the convenience of traditional hot-swap drives with a hardware RAID protection. At the same time, PCIe SSD adapters do not support hot-swap capabilities, and they can use operating system's software RAID capabilities to offer data protection if required.

The highest IOPS density and the lowest latency is provided by the eXFlash DIMMs that are highly optimized for both IOPS and latency. And, they can support the most demanding IOPS-intensive and latency-sensitive applications, including financial services, high-performance databases, big data analytics, and virtualization and cloud computing.

Technical overview

Lenovo eXFlash memory-channel storage (MCS) is the newest innovative flash storage technology that was introduced with X6 family servers. The eXFlash memory-channel storage is a high performance solid-state storage device that is produced in a standard DIMM form factor and plugs into existing memory DIMM slot. Figure 4 shows the eXFlash DIMM.

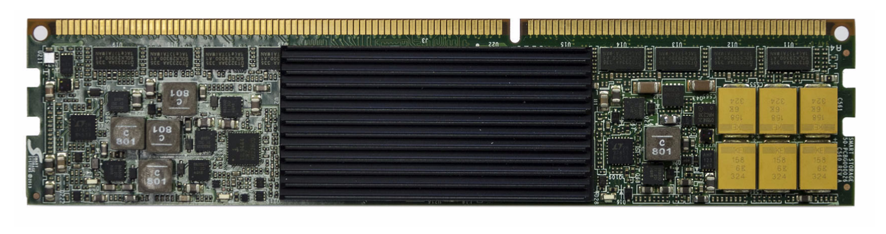


Figure 4 eXFlash DIMM

The eXFlash DIMM modules are installed into DDR3 slots and use memory channels of the processors. Data transfers between processors and eXFlash DIMMs run directly without any extra controllers such as PCIe controller and SAS/SATA controllers. This approach can significantly reduce latency and improve performance.

Figure 5 shows the difference in data access between eXFlash DIMMs and other solid-state storage products, such as PCIe SSD adapters and SAS or SATA SSDs.

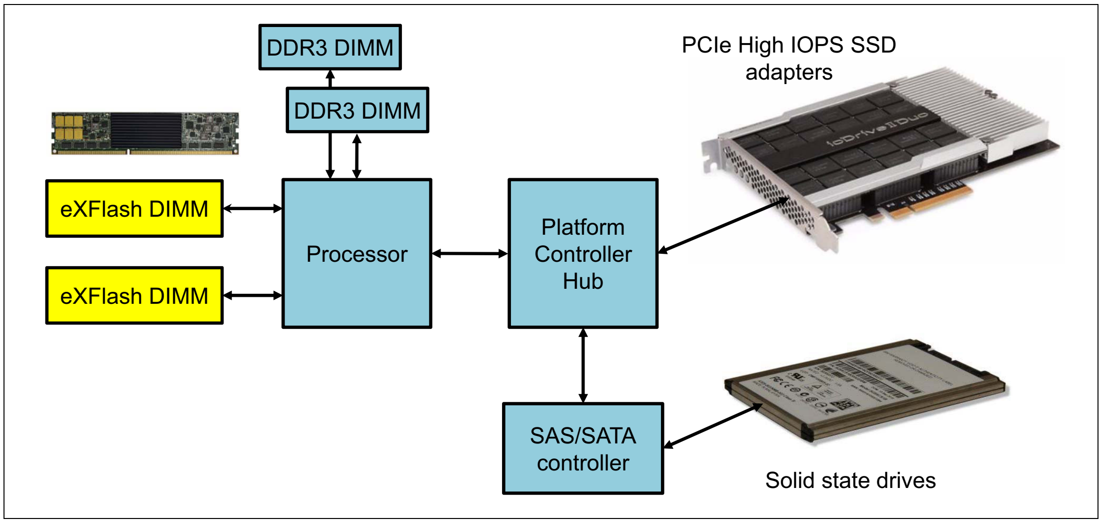


Figure 5 eXFlash DIMMs placement

Features and specifications

Key features of the eXFlash memory-channel storage are as follows:

- Ultralow write latency with WriteNow technology
  - Less than 5 microseconds response time
  - Less wait time between transactions
  - Deterministic response time across varying workloads
  - Tight standard deviation on performance
  - Consistent performance for highest throughput and speed
- High scalability
  - Add multiple eXFlash DIMMs without experiencing performance degradation
  - Highest flash storage density within the server
- Maximized storage footprint with utilization of existing unused DDR3 slots
  - Increases storage capacity without increasing your servers
  - Features industry-standard DDR3 form factor
  - Plugs into existing DDR3 slot

The eXFlash DIMM leverages cost-efficient consumer-grade 19 nm MLC flash with FlashGuard technology, which provides up to ten drive writes per day (DWPD) to meet the endurance needs of write-intensive and mixed-used application workloads.

eXFlash DIMMs have the following key characteristics:

- Provide ultrahigh overall performance using array of eXFlash DIMMs in a parallel manner.
- Up to 12.8 TB total flash storage capacity per server with 200 GB and 400 GB eXFlash DIMMs.

- Ultra low latency of less than 5 µs write latency.
- eXFlash DIMMs use available DDR3 memory channels; they support up to 1600 MHz DDR memory speeds.
- eXFlash DIMMs can be intermixed with standard registered memory DIMMs (RDIMMs) on the same memory channel.
- eXFlash DIMMs support advanced reliability and availability features including Flexible Redundant Array of Memory Elements (F.R.A.M.E.), DataGuard, and EverGuard.
- eXFlash memory-channel storage is supported by the major operating systems through software drivers

This new technology allows supported System x® servers to deliver breakthrough performance for targeted workloads by offering significantly lower latency compared to traditional solid-state drives like eXFlash SSDs, and even PCIe SSD adapters like High IOPS Adapters.

Table 3 summarizes specifications of eXFlash DIMM offerings.

Table 3 eXFlash DIMM specifications

Specification	**200 GB**	**400 GB**
Part number	00FE000	00FE005
Interface	DDR3 (800 - 1600 MHz)	DDR3 (800 - 1600 MHz)
Hot-swap	No	No
Form factor	LP DIMM	LP DIMM
Capacity	200 GB	400 GB
Endurance	10 drive writes per day (5 year lifetime expectancy)	10 drive writes per day (5 year lifetime expectancy)
Maximum power	12 W	12 W

Table 4 summarizes performance characteristics of eXFlash DIMM offerings.

Table 4 eXFlash DIMM specifications

Specification	**200 GB**			**400 GB**		
Part number	00FE000			00FE005		
Server family tested	System x3650 M4 (E5-2600 v2)		X6 servers	System x3650 M4 (E5-2600 v2)		X6 servers
Operational speed	1600 MHz	1333 MHz	1333 MHz	1600 MHz	1333 MHz	1333 MHz
Maximum read IOPS[a]	135,402	135,525	144,672	135,660	135,722	139,710
Maximum write IOPS[a]	28,016	28,294	29,054	41,424	41,553	43,430
Maximum sequential read rate[b]	743 MBps	689 MBps	644 MBps	739 MBps	696 MBps	636 MBps
Maximum sequential write rate[b]	375 MBps	376 MBps	382 MBps	388 MBps	392 MBps	404 MBps
Read latency[c]	150 µs	151 µs	141 µs	150 µs	151 µs	144 µs
SEWC write latency	4.66 µs	5.16 µs	6.78 µs	4.67 µs	5.17 µs	7.08 µs

[a] 4 KB block transfers

b. 64 KB block transfers
c. Latency measured at hardware (CLAT) exclusive of system latency (SLAT).

For more information, see the Product Guide, *eXFlash DDR3 Storage DIMMs*, TIPS1141, available from:

http://www.redbooks.ibm.com/abstracts/tips1141.html

Architecture and components

The eXFlash DDR3 Storage DIMM has the following onboard components:

- 19 nm MLC NAND flash memory modules
- Flash controllers which implement advanced flash management and protection techniques
- Memory controller chipset, which provides an interface between physical DDR3 memory bus and solid-state storage
- Power system, which protects memory write buffers from the unexpected power outage

The eXFlash DIMM hardware components are shown in Figure 6.

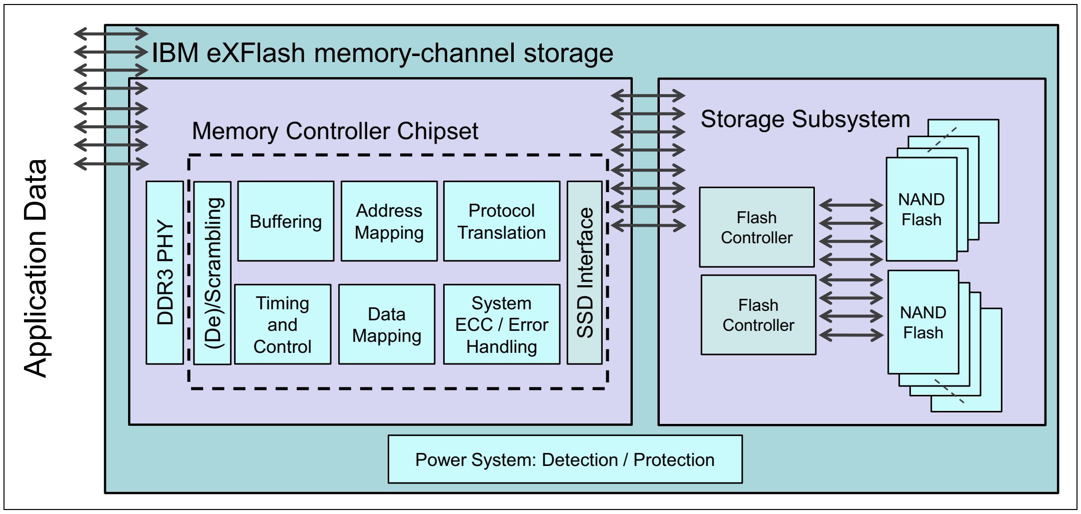


Figure 6 eXFlash DIMM block diagram

eXFlash DIMMs support the following advanced flash management technologies:

- FlashGuard includes innovative technologies to reliably extract significantly more usable life from the traditional consumer-grade MLC flash than provided by the standard specifications published by NAND manufacturers.
- DataGuard provides full data path protection ensuring that user data will be safe throughout the entire data path, and it provides the ability to recover data from failed page and NAND blocks.
- EverGuard prevents the loss of user data during unexpected power interruptions.

FlashGuard incorporates two important technology breakthroughs in the area of flash management:

- Aggregated Flash Management technology prolongs the life of solid-state devices by treating all flash elements in them as a system instead of as a collection of discrete elements. Aggregating the management of the flash over multiple pages within a block and over multiple blocks within the solid-state device reduces the limitations imposed at the page and block levels, thus extending the useful life of the drive.

- Advanced Signal Processing technology is used to continually monitor the flash modules and collect detailed statistics of their performance. This information is used to dynamically adjust the flash operating parameters to attain maximum endurance from the drive throughout its operational life.

DataGuard technology protects all user data while it is being transferred to or from the flash memory and while it resides within the drive.

This full data path protection ensures that user data will be safe from undetected electrical and firmware failures across all points in the drive. DataGuard also includes a cross-die data redundancy feature called Flexible Redundant Array of Memory Elements (F.R.A.M.E.) that enables the recovery of user data in the event of catastrophic events such as flash page or block failures.

EverGuard technology includes an array of highly reliable solid-state capacitors that act as an independent power supply for the drive during power interruptions and ensure that all user data is transferred from the write cache to the flash memory.

eXFlash DIMMs are recognized by the server as solid-state storage devices like many other block storage devices. The specialized kernel driver is required for the operating system to use eXFlash DIMMs.

The Lenovo eXFlash channel-memory storage system architecture and operations are shown in Figure 7.

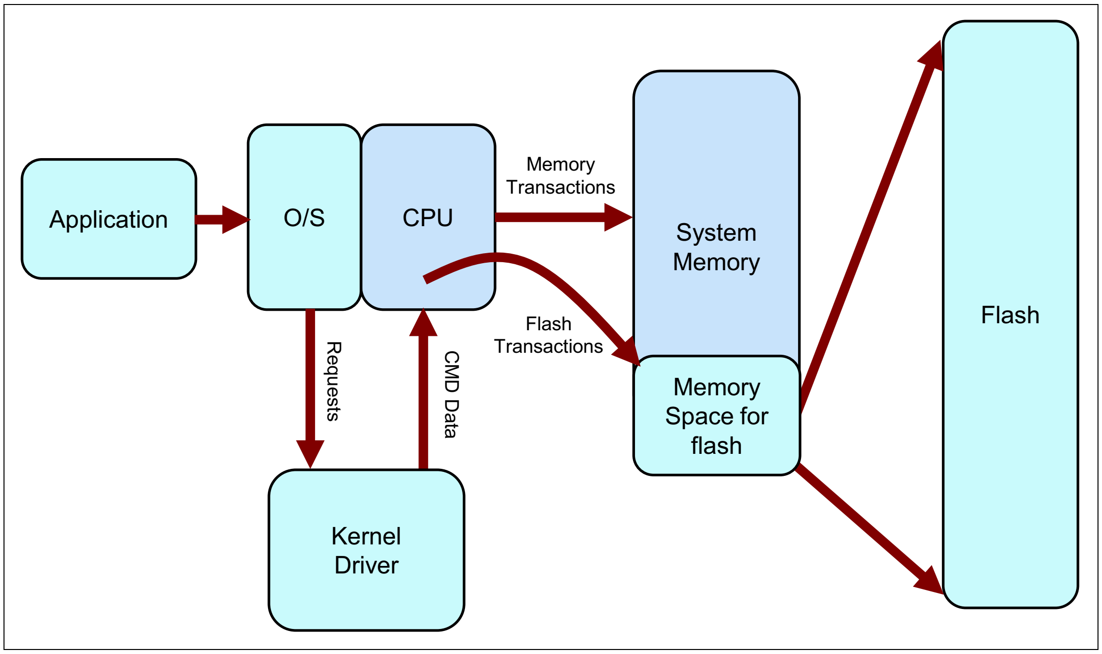


Figure 7 Lenovo eXFlash channel-memory storage system architecture and operations

Traditional DRAM memory and eXFlash DIMM memory address spaces are logically isolated from each other by the server's UEFI firmware. Some system memory DRAM space is reserved for eXFlash DIMM operations.

When the application sends the storage I/O request to the operating system (OS), the OS forwards it to the eXFlash DIMM kernel driver, which generates respective memory-channel commands to access the data stored on the eXFlash DIMMs. The requested data is then transferred from the flash memory to the system memory directly via a DDR3 memory bus.

Supported servers and operating systems

For a complete list of supported servers and operating systems, including configuration rules, see the Product Guide, *eXFlash DDR3 Storage DIMMs*, TIPS1141, available from:

http://www.redbooks.ibm.com/abstracts/tips1141.html?Open

For more information, refer to *eXFlash DIMM Configuration and Support Requirements:*

http://ibm.com/support/entry/portal/docdisplay?lndocid=SERV-FLASHDM

Use cases

This subsection offers ideas about where to deploy Lenovo eXFlash memory-channel storage alone or combined with other Lenovo offerings, such as FlashCache Storage Accelerator™, and their potential benefits in these solutions. The following scenarios are described:

- "Transactional databases (OLTP)"
- "Data warehouses (OLAP)" on page 17
- "Virtualized workloads" on page 19

Transactional databases (OLTP)

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.

OLTP workloads are characterized by small, interactive transactions that generally require subsecond response times. The key performance indicator (KPI) of the transactional system is latency, because the user expects to receive the requested product information or to place an order quickly. Inability to meet these user expectations leads to customer dissatisfaction and revenue loss. eXFlash memory-channel storage addresses these challenges by providing low latency, extreme performance, and efficient transaction management.

There are two ways to implement eXFlash memory-channel storage in OLTP solutions:

- eXFlash DIMMs as caching devices
- eXFlash DIMMs as a primary data storage

When eXFlash DIMMs are used for caching, FlashCache Storage Accelerator is used. FlashCache Storage Accelerator is a suite of caching software and tools that are designed to significantly increase your server and storage performance and efficiency by helping eliminate I/O bottlenecks, and keep the most active data closer to the application. This in turn helps you gain greater system utilization, lower software and hardware costs, and save power, cooling, and floor space.

Used in conjunction with flash storage offerings, FlashCache Storage Accelerator can provide an efficient, cost-effective, and easy-to-implement solution for virtual and physical environments.

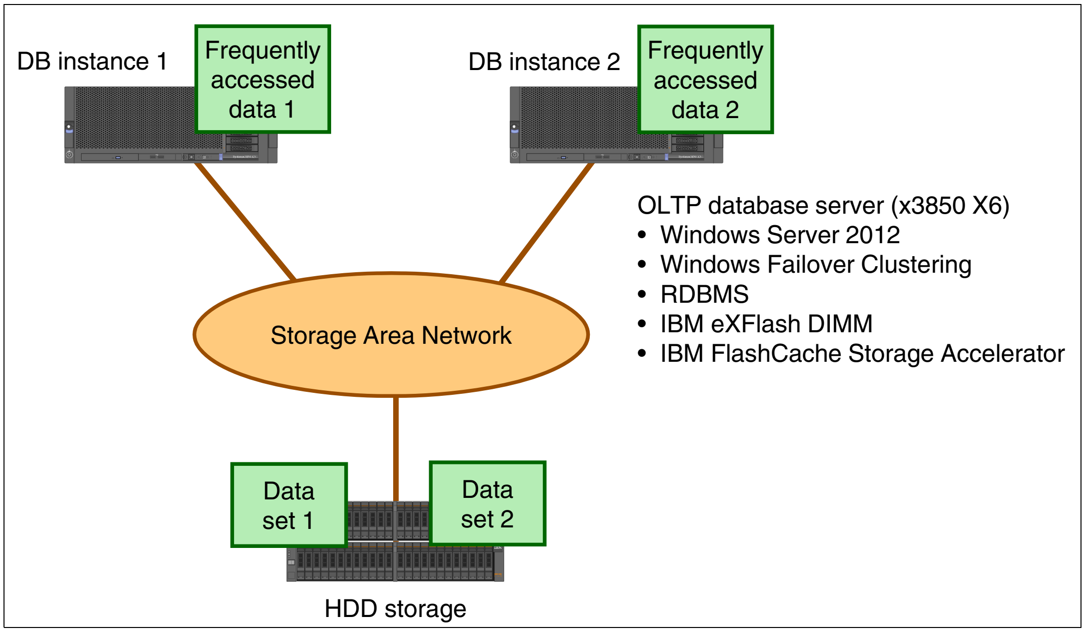
FlashCache Storage Accelerator includes the following key features:

- Transforms flash storage (SSDs, High IOPS SSD Adapters, and eXFlash) into a transparent acceleration device to cache frequently accessed data off of any storage system, whether SAN or DAS.
- Caches data within servers, close to the application workloads, delivering extremely low latency.
- Works in physical (non-virtualized) and virtual environments.
- Transparently and dynamically rebalances I/O.
- Includes disk, file, volume and VM-specific caching.
- Is supported on a wide variety of operating systems
- Supports both native operating system high availability clustering (such as Microsoft Failover Clustering and Linux high availability clusters) and advanced workload availability and re-balancing for virtualized environments (VMware vMotion, HA, FT, and DRS).

An OLTP solution with eXFlash DIMMs and FlashCache Storage Accelerator typically consists of the following components:

- System x database servers (typically, X6 systems) that run data management software such as IBM DB2, Microsoft SQL Server, or Oracle Database. These database servers also host the following items:
  - FlashCache Storage Accelerator Direct edition for Windows
  - eXFlash DIMM
- External shared storage system (such as IBM Storwize V3700) that host the entire data set.
- Storage area network (SAN) that is used to provide connectivity across database servers and storage systems.

OLTP database acceleration scenario is shown in Figure 8.



OLTP database server (x3850 X6)
- Windows Server 2012
- Windows Failover Clustering
- RDBMS
- IBM eXFlash DIMM
- IBM FlashCache Storage Accelerator

Figure 8 OLTP database acceleration

System x servers are developed, tested for quality, and certified by Lenovo, and they are backed by incomparable worldwide service and support from Lenovo and IBM. System x servers deliver business value over the long term because of advancements in scalability, reliability, and performance. These advancements are combined with flexible configuration options, energy efficient components, and robust systems management tools.

X6 servers, such as System x3850 X6, are designed for mission-critical enterprise-class workloads such as databases. By using open, industry-standard components, which are combined with X6 technologies, these systems provide leadership performance, scalability, and reliability.

In this scenario, database server systems are connected to the external shared storage through Fibre Channel SAN. Host systems run Windows Server 2012 operating system. They support multipathing to allow redundant storage connections through SAN, including dual-port storage system interface connections, dual-port host bus adapters (HBAs) on the host systems, and redundant SAN switched fabric.

Database high availability is achieved by using the Windows Failover Clustering service, and two database instances are running in the cluster. The entire data sets for the database instances (Data set 1 and Data set 2) are hosted on the external SAN-attached storage, and server-side caching is implemented with the FlashCache Storage Accelerator and eXFlash DIMMs.

The most frequently used data from the data sets on the SAN are dynamically and transparently cached on the server's local eXFlash memory-channel storage to increase performance and lower response time.

In the second scenario, the entire data set is placed onto an eXFlash DIMM internal storage in the server. eXFlash DIMMs are configured in mirrored pairs to provide data redundancy.

In addition, to ensure that the high availability requirements are met in case of a node failure, the following techniques can be used depending on the database vendor:

- Log shipping
- Replication
- Database mirroring

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

If, for some reason, the entire database cannot be placed onto eXFlash, consider putting at least part of the data there. Look at the following areas:

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

Data warehouses (OLAP)

Data warehouses are commonly used with online analytical processing (OLAP) workloads in decision support systems, for example, financial analysis. Unlike OLTP, where transactions are typically relatively simple and deal with small amounts of data, OLAP queries are much more complex and process large volumes of data. By its nature, the OLAP workload is sequential read-intensive and throughput-intensive.

OLAP databases are normally separated from OLTP databases; OLAP databases consolidate historical and reference information from multiple sources. Queries are submitted to OLAP databases to analyze consolidated data from different perspectives to make better business decisions in a timely manner.

Impact of slow storage I/O in OLAP environments

For OLAP workloads, a fast response time is critical to ensure that strategic business decisions can be made quickly in dynamic market conditions. Delays can significantly increase business and financial risks. Usually, decision-making is stalled or delayed because of a lack of accurate, real-time operational data for analytics. This means missed opportunities for the following reasons:

- Inability to gain insight into a business
- Inability to predict business outcomes
- Explosion of volume, variety, and velocity of information

The delays are primarily from batch data loads and performance issues because of handling heavy complex queries that use I/O resources. A common performance bottleneck in OLAP environments is the I/O that is required for reading massive amounts of data (frequently referred to as *big data*) from storage for processing in the OLAP database server. The server ability to process this data is usually a non-factor because the server typically has significant amounts of RAM and processing power, parallelizing tasks across the computing resources of the servers.

In general, clients might experience the following challenges in OLAP environments:

- Slow query execution and response times, which delays business decision making
- Dramatic growth in data, which requires deeper analysis

Lenovo eXFlash memory-channel storage solution for OLAP

eXFlash memory-channel storage can help make businesses more agile and analytics-driven by providing up-to-the-minute analytics based on real-time data, and not yesterday’s news, in the following ways:

- Dramatically boosting the performance of OLAP workloads with distributed scale-out architecture, providing almost linear and virtually unlimited performance and capacity scalability
- Significantly improving response time for better and timely decision-making

Flash storage acceleration solutions for OLAP use a distributed server and storage scale-out approach. This approach satisfies high bandwidth requirements and provides unlimited performance and capacity growth capabilities, matching the growing volumes of data that is being processed.

There are two ways to use eXFlash DIMM storage in OLAP solutions:

- eXFlash DIMM as primary data storage
- eXFlash DIMM as a caching device

eXFlash DIMM as primary data storage

An OLAP solution with eXFlash DIMM as primary data storage (shown in Figure 9 on page 18) consists of the following components:

- System x database servers (typically, dual-socket x3650 M4 systems) that run data management software. These database servers also host the following item:
  - eXFlash DIMMs for the partitioned data sets
- A private network (such as 10 Gb Ethernet or QDR/FDR InfiniBand) that is used to provide high-speed connectivity across database servers in a cluster.

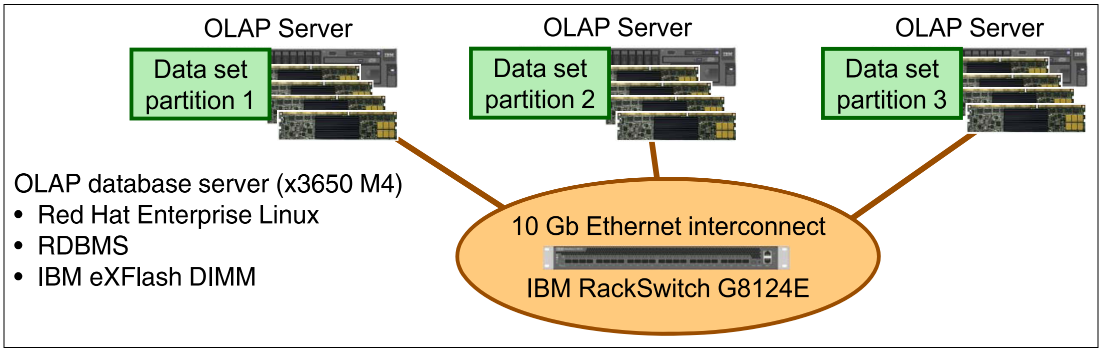


Figure 9 OLAP database acceleration

Server hosts, or nodes, are interconnected with the isolated high-speed network (such as 10 Gb Ethernet with RackSwitch™ G8124E) that is used for the inter-node data exchange. Each node runs a copy of the OLAP database application, and the analyzed data set is partitioned and distributed across the storage systems. Depending on the database management software that is used and its architecture, each node might have access to only a certain portion of data. OLAP queries are distributed and processed across the nodes.

This solution can scale easily by adding more similarly configured nodes. In such a case, storage capacity and I/O bandwidth are incremented linearly with the increasing number of nodes, which can help to eliminate storage I/O bottlenecks in OLAP workloads.

eXFlash DIMM as a caching device

An OLAP solution with eXFlash DIMM as a caching device is based on the same architecture as the previously described solution (“eXFlash DIMM as primary data storage” on page 18). In addition, it includes FlashCache Storage Accelerator software. The solution that is shown in Figure 10 consists of the following components:

- System x database servers (typically, dual-socket x3650 M4 systems) that run data management software. These database servers also host these items:
  - FlashCache Storage Accelerator Direct edition for Linux
  - eXFlash DIMMs
  - Local HDD storage for the partitioned sets of the data that is being analyzed
- A private network (such as 10 Gb Ethernet or QDR/FDR InfiniBand) that is used to provide high-speed connectivity across database servers in a cluster.

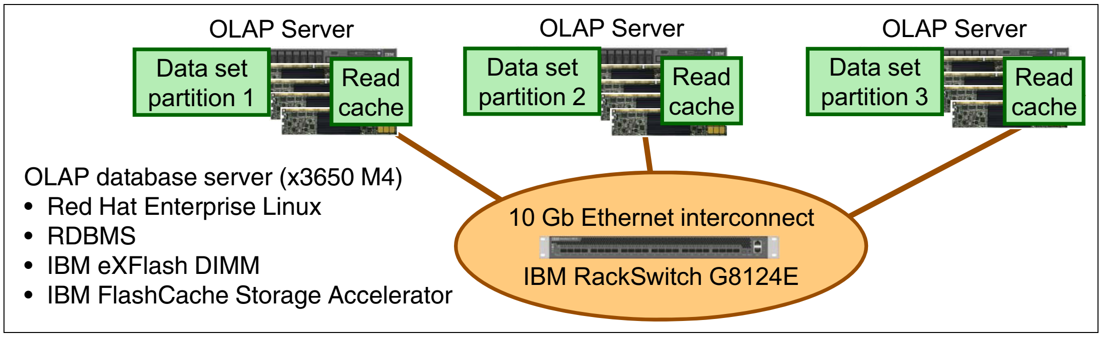


Figure 10 OLAP database acceleration

Virtualized workloads

Although many users realize benefits on savings from deploying server virtualization solutions, some questions and concerns should still be addressed to move the adoption of virtualization to the next level. These include applications that could not be virtualized before because of existing storage I/O constraints.

One of the key requirements of any virtualized environment is high availability for virtual machines in addition to easy VM migration. However, most high availability technologies require shared access to storage from multiple physical servers to enable the migration of VMs across hosts.

Traditional data center architectures rely on expensive SAN-based storage systems for shared storage access. However, SAN storage can become a performance bottleneck when deployed with clusters of VM-dense servers, and is expensive to scale. The limited IOPS available for a SAN storage array must be shared with many VMs across multiple physical servers. Additionally, there is no easy way to segregate I/O traffic on a per-VM basis to allocate IOPS to the specific VMs that need them.

However, the overall architecture still needs to maintain shared storage access across physical servers to enable VM mobility. Although, it is possible to use local flash in physical servers as static cache, this approach ties VMs to physical servers and breaks VM mobility. Furthermore, in a mixed virtualized environment, various kinds of workloads and storage access patterns, even on the same physical server, can be observed.

eXFlash memory-channel storage, combined with FlashCache Storage Accelerator, can help solve many challenges of a virtualized environment by transparently and dynamically adapting to the changing workload patterns while keeping essential VM mobility and availability features and increasing storage IOPS performance.

In virtual environments, the FlashCache Storage Accelerator is deployed as a part of the hypervisor (VMware ESXi), as shown in Figure 11 on page 20. Optionally, for more granular cache management, FlashCache Storage Accelerator elements can be installed within the guest operation system, in addition to the hypervisor installation.

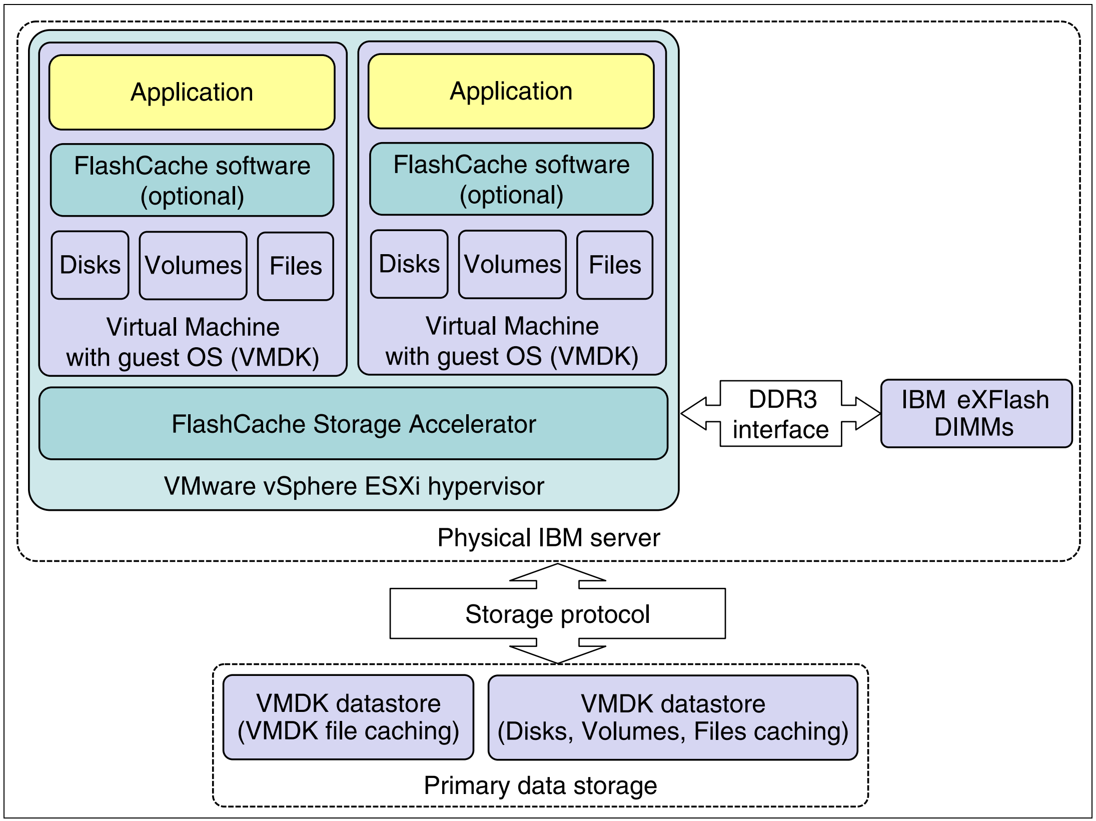

Application
FlashCache software (optional)
Disks
Volumes
Files
Virtual Machine with guest OS (VMDK)
Application
FlashCache software (optional)
Disks
Volumes
Files
Virtual Machine with guest OS (VMDK)
FlashCache Storage Accelerator
VMware vSphere ESXi hypervisor
DDR3 interface
IBM eXFlash DIMMs
Physical IBM server
Storage protocol
VMDK datastore (VMDK file caching)
VMDK datastore (Disks, Volumes, Files caching)
Primary data storage

Figure 11 FlashCache Storage Accelerator in virtual environments

The following caching levels are supported in the virtual environments when the FlashCache Storage Accelerator components are installed on the VMware ESXi hypervisor and within the guest operating system (OS):

- Disk caching
- Volume caching
- File caching
- VMDK file caching

If the FlashCache Storage Accelerator is installed only as a part of the hypervisor, only VMDK file caching is available.

Performance consideration: As a general rule, better performance results can be achieved when the FlashCache Storage Accelerator is deployed within the guest OS.

The FlashCache Storage Accelerator supports storage-specific VMware multipathing software to take full advantage of redundant SAN storage connections.

The FlashCache Storage Accelerator transparently supports VMware high availability clustering (VMware HA), live migration (vMotion), and dynamic resource reallocation (Distributed Resource Scheduling, or DRS).

The management of the FlashCache Storage Accelerator environment is performed through the Flash Management Console. The management console should be installed on a separate non-cached virtualized server, and all instances of the FlashCache Storage Accelerator can be managed remotely from this console.

Performance benchmarking

The Lenovo Performance Lab conducted performance benchmarks for the eXFlash DIMMs to validate their scaling capabilities within a single system and compare them to the conventional solid-state drives (SATA Enterprise SSDs).

Figure 12 highlights scaling capabilities, comparing relative random read and write IOPS performance for one and four eXFlash DIMMs (higher value is better, as the figure shows).

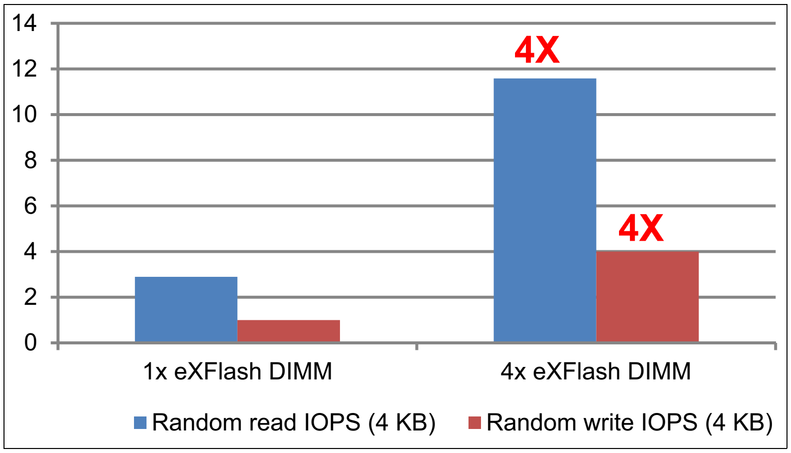


Figure 12 eXFlash DIMM scalability

As shown in Figure 12, eXFlash DIMMs provides near linear performance scalability by adding more eXFlash DIMMs to the system.

Figure 13 compares relative IOPS performance of four eXFlash DIMMs versus four Enterprise SATA SSDs (higher is better).

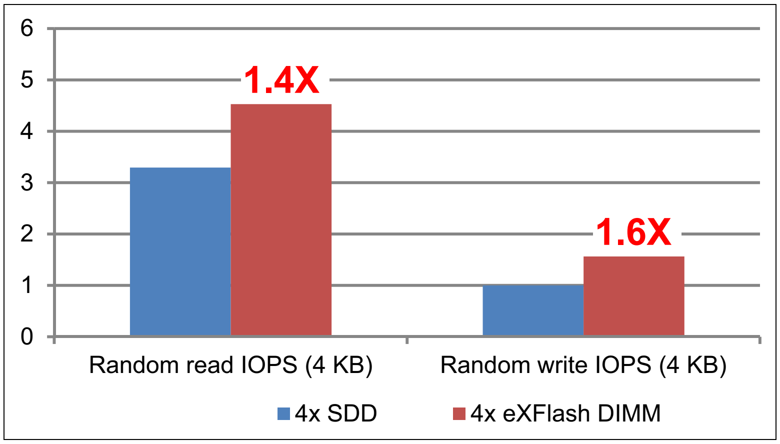


Figure 13 eXFlash DIMM versus SSD: IOPS performance

The eXFlash DIMMs provides up to 40% better read IOPS performance, and they also improve write IOPS performance by almost 60%.

The next two charts compare relative read latency (Figure 14) and write latency (Figure 15 on page 22) for the eXFlash DIMMs versus SSDs (lower is better).

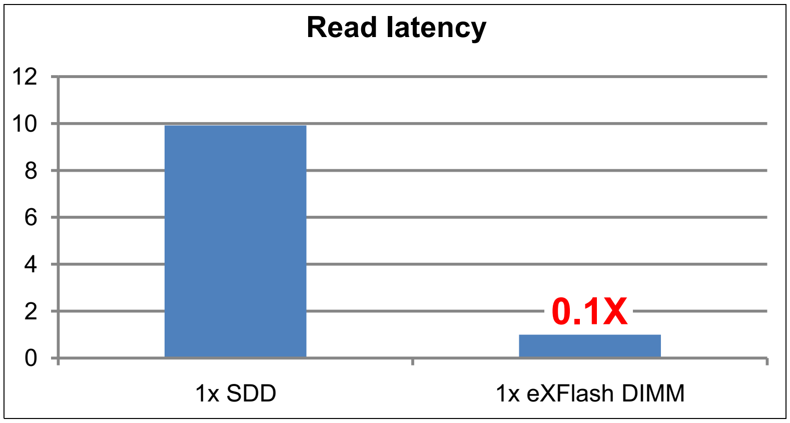


Figure 14 eXFlash DIMM versus SSD: read latency

With queue depth of 64, the eXFlash DIMM provides more than 10 times improvement in read access latency compared to an SSD.

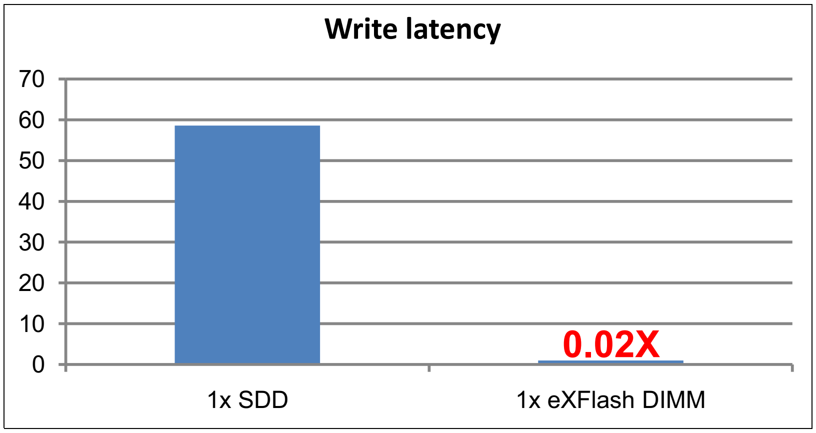


Figure 15 eXFlash DIMM versus SSD: write latency

With a queue depth of 1, the eXFlash DIMM provides more than 50 times improvement in write access latency compared to an SSD.

Conclusion

The growth in CPU processing power far exceeds the growth in storage I/O. For this reason, storage I/O is the culprit in major bottlenecks in many performance-demanding applications. The eXFlash memory-channel storage can significantly improve storage I/O throughput, making it possible for you to potentially eliminate I/O bottlenecks in a system.

eXFlash memory-channel storage offers the highest IOPS density per GB and the lowest write latency among flash storage offerings for the most demanding IOPS-intensive and latency-sensitive applications. With eXFlash DIMMs, you can increase data access speed up to 10 times compared to conventional SSDs.

eXFlash memory-channel storage, combined with FlashCache Storage Accelerator, can help virtualize data-intensive enterprise applications that were unable to be virtualized because of storage I/O constraints. This provides higher reliability and availability of services, lower acquisition costs, and shortens ROI time frame and decreases overall TCO.

The combination of eXFlash DIMMs and Storage Accelerator can also help boost the application performance in physical (non-virtualized) Windows and Linux environments, and it provides a single point of management for both virtualized and non-virtualized systems.

Related publications

The following Lenovo Press publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *eXFlash DDR3 Storage DIMMs Product Guide,* TIPS1141
- *System x3950 X6 Product Guide,* TIPS1132
- *X6 Servers: Technology Overview,* REDP-5059
- *Workload Optimization with X6 Servers,* REDP-5058
- *The Benefits of FlashCache Storage Accelerator in Enterprise Solutions,* REDP-5080
- *System x3850 X6 Planning and Implementation Guide,* SG24-8208

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, drafts, and additional materials at the following website:

ibm.com/redbooks

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO), Raleigh Center.

Ilya Krutov is a Lenovo Press Project Leader in the Enterprise Business Group Brand Enablement team. He manages and produces pre-sale and post-sale technical publications for various IT topics, including x86 rack and blade servers, server operating systems and software, virtualization and cloud, and datacenter networking. Ilya has more than 15-year experience in the IT industry, and he has been performing various roles, including Run Rate Team Leader, Portfolio Manager, Brand Manager, IT Specialist, and Certified Instructor. He has written more than 180 books, papers and other technical documents. He has a Bachelor's degree in Computer Engineering from the Moscow Engineering and Physics Institute (Technical University).

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

From Lenovo:

- Jacqueline Gutierrez
- Joe Jakubowski
- David Watts

From IBM Redbooks:

- Tamikia Barrow
- Cheryl Gera
- Chris Rayns
- Diane Sherman
- Debbie Willmschen

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consulty our local Lenovo representative for information on the products and services currently available in yourarea. 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.

This document REDP-5089-00 was created or updated on December 8, 2014.

Send us your comments in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an email to:
  redbooks@us.ibm.com

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™
FlashCache™
FlashCache Storage Accelerator™
Lenovo®
RackSwitch™
Lenovo(logo)®
System x®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, 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.