High Performance Solution for Oracle Database 12c with Lenovo x3650 M5 Server and SanDisk Fusion ioMemory SX350 Storage

Examines how to reduce TCO by running Oracle workloads more efficiently

Provides a deployable solution for accelerating Oracle database workloads

Examines flash storage based server performance of OLTP and DSS workloads

Evaluates how employing flash storage dramatically reduces completion of database I/O requests

Mark Johnson, SanDisk

Prasad Venkatachar, Lenovo

Ron Kunkel, Lenovo

LENOVO PRESS

Abstract

Oracle users often say their top priority is reliability, followed by a blend of performance, capacity, and cost. These requirements provide the framework for this paper. To reduce costs and achieve a rapid return on investment, enterprises can run large performance-sensitive Oracle databases with confidence, or consolidate databases from many servers. Leveraging flash storage allows I/O requests to complete faster by an order of magnitude over traditional disk drive storage.

This paper provides a set of deployable solutions for accelerating various Oracle Database workloads with SanDisk PCIe-based flash storage while simultaneously reducing TCO. Two database workloads are examined in this paper: On-Line Transaction Processing (OLTP), and Decision Support System (DSS).

As shown in this paper, using SanDisk PCIe-based flash storage all but eliminates the processor's spin-wait events and allows the system as a whole to perform significantly more operations per unit of time.

At Lenovo® Press, we bring together experts and companies to produce technical publications for topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges. In this case, Lenovo collaborated with SanDisk and Oracle to bring you this paper.

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

`http://lenovopress.com`

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Executive summary

Running on Oracle Database 12c, the Lenovo x3650 M5 server provides an affordable, small-footprint enterprise solution that delivers both high reliability and performance. To demonstrate solution benefits we have employed the HammerDB tool for Oracle testing. With the Oracle workload this solution delivers over a million database IOPS (8KB block size) with submillisecond latency, for over 2 million HammerDB OLTP transactions per minute. The storage layer also exceeds 16GB/s bandwidth for data warehouse and analytics workloads.

To reduce costs and achieve a rapid return on investment, enterprises can run large performance-sensitive Oracle databases with confidence, or consolidate databases from many servers.

Oracle users often say that their top priority is reliability, followed by a blend of performance, capacity, and cost. These requirements provide the framework for this paper. The foundation of the solution starts with the x3650 M5, which was configured with dual 14-core Intel Xeon E5-2680 v4 processors to provide a perfect balance of performance and cost. For the storage layer, the SanDisk Fusion ioMemory SX350 Application Accelerator was selected. This device offers flash storage capacities up to 6.4 TB per storage card. The combined server and storage elements deliver high performance on both read and write workloads.

This paper provides a set of deployable solutions for accelerating various Oracle database workloads while simultaneously reducing TCO. We examine two database workloads in this paper:

- On-Line Transaction Processing (OLTP)

  OLTP is characterized by a large set of users issuing small transactions with approximately 60-70% reads and 30-40% writes. Therefore, the underlying storage must offer high concurrency, high endurance, high IOPS, low latency, and the ability to sustain heavy writes from the Oracle Database Writer, Log Writer, and Archiver processes.

- Decision Support System (DSS)

  DSS, warehouse, and analytic workloads are all characterized by bulk-loading vast amounts of data, followed by countless long sequential reads. That means the storage layer must offer high capacity and high bandwidth on both reads and writes, but not necessarily the same level of concurrency or endurance as OLTP.

It is important to design and set up a fully optimized system to deliver a high performance Oracle solution. The x3650 M5 supports up to 1.5 TB of DDR4 memory. For the testing conducted to support this paper, the x3650 M5 server was configured with 256 GB of memory and the size of the dataset under test was 5 TB. For datasets that exceed the provisioned server's DRAM capacity, as they were in this testing, the server's ability to support an application like Oracle Database 12c will be restrained by the I/O performance of the server's storage. The Fusion ioMemory SX350 storage employed in this testing shows how flash storage can be used to boost overall system performance.

Each storage card offers 6.4 TB of flash storage, and the x3650 M5 accommodates six storage cards for a total capacity of 38.4 TB. The Fusion ioMemory SX350 offers excellent performance on both reads and writes, which is unique among flash storage products.

Performance experts and Oracle database administrators rely on Oracle AWR reports to identify the bottlenecks in their database systems. If Oracle AWR reports indicate that the servers' CPUs are not constantly engaged, then the storage system is not effectively supporting the application's I/O demands. Slow I/O response of traditional storage systems using hard disk drives can significantly degrade application performance.

Using flash storage allows I/O requests to complete quicker by orders of magnitude. This all but eliminates the processor's spin-wait events and allows the system as a whole to perform significantly more operations per unit of time. In many cases, the implementation of SanDisk flash storage allows the database to run faster by more effectively serving the data needs of the database application. This x3650 M5-based solution yields a well balanced system of compute, memory and storage resources to deliver superior Oracle Database 12c performance.

This paper is intended for IT professionals running Oracle databases and should be of interest to system, storage, and database administrators alike.

Solution value: Lower TCO

Oracle users want reliability without sacrificing performance or increasing cost. Many of the high-performance solutions available today are cost-prohibitive to acquire and complex to maintain. The goal of this effort was to design a simple and affordable solution that satisfies the requirements of demanding Oracle users.

The Lenovo System x3650 M5 server with Fusion ioMemory SX350 storage cards provided a reliable, high-performance solution for running Oracle workloads that efficiently used system resources to help drive down total cost of ownership (TCO).

To derive a solid return on investment for Oracle solutions it is important to efficiently utilize all the processors in the system. Right-sizing the number of processors for a given application requirement will help to reduce the cost. Another option for reducing TCO is to consolidate multiple databases from different servers onto a single x3650 M5 server and still meet high-performance requirements.

It is not unusual for Oracle users to deploy databases on separate servers in order to manage CPU capacity. However, this typically results in deployments that feature a collection of underutilized servers, which increases capital and operating costs. One reason for this deployment practice is a misunderstanding of CPU utilization in database servers. The CPU may look busy, but in fact it is waiting for data to be fetched by slow storage devices.

Moving the database to a modern storage platform allows the I/O calls to complete orders of magnitude faster; thereby reducing the CPU waits and enabling more efficient use of the CPU. Using the consolidated approach outlined in this paper, an Oracle 12c Database customer can eliminate the number of older-generation servers. This consolidated server approach reduces capital infrastructure costs and lowers operating expenses.

Server consolidation is a great way to reduce operating costs. There are two basic consolidation models to consider:

- Running multiple, physically separate databases per server
- Using Oracle Multitenant introduced in Oracle Database 12c

Running multiple databases on a single server can certainly be achieved without deploying and enabling Oracle Multitenant, but the databases would have redundancies, such as memory buffers, temporary and undo tablespaces and the redo and archive log files. Oracle Multitenant eliminates these redundancies to more efficiently use server resources. Thus, the Oracle Database 12c user can achieve maximum efficiency and savings by deploying Oracle Multitenant on flash-powered x3650 M5 servers with Fusion ioMemory SX350 storage.

Featured products

The following sections describe the hardware products featured in this paper: the Lenovo System x3650 M5 server and Fusion ioMemory SX350 storage.

The Lenovo System x3650 M5 server

The Lenovo System x3650 M5, Figure 1, is a versatile 2U dual-socket, business-critical server that offers improved performance and pay-as-you-grow flexibility, along with new features that improve server management capability. This powerful system is designed for an enterprise's most important business applications and cloud deployments.

Figure 1 Lenovo System x3650 M5

With the powerful, versatile 2U rack server design, the dual-socket x3650 M5 server can run a wide array of workloads, 24x7, helping enterprises gain faster business insights. Incorporating the Intel Xeon E5-2600 v4 processor product family and industry-leading storage capacity for a two-socket server, the x3650 M5 delivers exceptional performance. Flexible and scalable internal storage configurations include up to 28 2.5-inch drives.

Combining balanced performance and flexibility, the x3650 M5 is a great choice for small business, medium businesses and large enterprises. It provides outstanding uptime to keep business-critical applications and cloud deployments running safely and reliably. Ease of use and comprehensive systems management tools reduce deployment complexity. Extreme reliability, availability, serviceability (RAS), combined with a high-efficiency design, improve business environments and help save operational costs.

SanDisk Fusion ioMemory SX350 Application Accelerator

The SanDisk Fusion ioMemory SX350 Application Accelerator Add-In-Card (AIC) combines VSL (Virtual Storage Layer) software with enterprise-grade Fusion ioMemory hardware to take enterprise applications and databases to the next level. The Fusion ioMemory SX350 provides consistent microsecond data access latency for mixed workloads, multiple gigabytes per second access, and hundreds of thousands of IOPS from a single product.

With industry-leading reliability (e.g., uncorrectable bit error ratio of only 1 in every 10^{20} bits read), the sophisticated Fusion ioMemory SX350 architecture allows for nearly symmetrical read and write performance with best-in-class low queue depth performance. This makes the Fusion ioMemory SX350 storage card ideal for a wide variety of real world, high-performance enterprise environments.

Figure 2 shows the 6.4TB SanDisk Fusion ioMemory SX350 Application Accelerator storage card (SX350-6400).

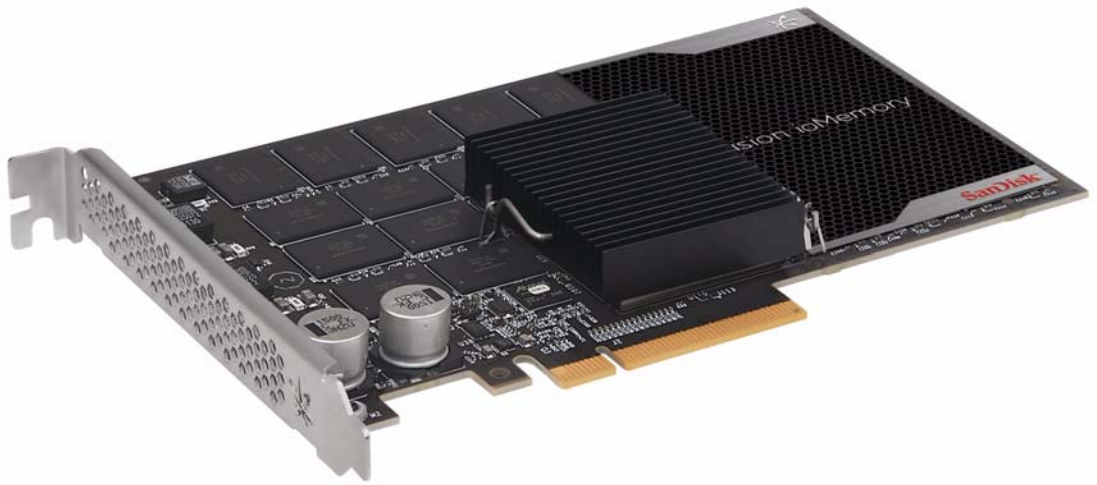

Figure 2 SanDisk Fusion ioMemory SX350 Application Accelerator storage card

This paper features the Fusion ioMemory SX350 model SX350-6400. Each storage card has 6.4TB raw usable capacity of SanDisk NAND flash memory. The x3650 M5 server was configured with six of these storage cards. The Fusion ioMemory SX350 storage card is available from Lenovo in per-device capacities ranging from 1.25TB to 6.4TB, with an endurance rating of 22 petabytes written. Table 1 lists the Lenovo part numbers.

Table 1 Lenovo part numbers for the Fusion ioMemory SX350 storage cards

Part number	Feature code	Description
00YA800	AT7N	io3 1.25TB Enterprise Mainstream Flash Adapter
00YA803	AT7P	io3 1.6TB Enterprise Mainstream Flash Adapter
00YA806	AT7Q	io3 3.2TB Enterprise Mainstream Flash Adapter
00YA809	AT7R	io3 6.4TB Enterprise Mainstream Flash Adapter

The Fusion ioMemory SX350 storage card requires only a PCIe 2.0 x8 slot, making it compatible with nearly all enterprise-class servers, and keeping data center costs to a minimum by consuming less than 25 watts of power under normal operating conditions. For this paper, the storage cards were configured to use no more than 48 watts under the heaviest of workloads.

Fusion ioMemory SX350 storage cards are unique in their ability to sustain high-capacity writes as well as reads. Most Oracle databases perform more reads than writes, but writes can still be a bottleneck for Oracle databases. Consider that many OLTP and Operational Data Stores have a workload consisting of 40% writes. These writes include inserts, updates, and deletes to row data and indexes, and the corresponding Undo, Redo, and Archive Log writes generated by these operations. Even Decision Support Systems (DSS) and databases used for On-Line Analytic Processing (OLAP) issue heavy writes on checkpoints, logging, and while ingesting bulk data. When selecting a storage product, it is imperative to consider the database's dependency on write operations not just read acceleration.

For more information about the SanDisk Fusion ioMemory SX350 Application Accelerator Add-in-Card described in this paper, and other SanDisk-branded flash storage devices, please visit https://www.sandisk.com/business/datacenter/products and http://lenovo.sandisk.com.

Architecture

This solution uses a single x3650 M5 server with dual Intel Xeon E5-2680 v4 14-core processors, 256GB RAM, and internal storage devices. No external storage is used, which greatly simplifies the solution, reduces administrative burden, and lowers TCO. The storage devices are Fusion ioMemory SX350 storage cards. These storage cards are detailed above in the section titled: SanDisk Fusion ioMemory SX350 Application Accelerator.

The following software was installed on the Oracle Database server:

- Oracle Linux 7.2 with Unbreakable Enterprise Kernel (UEK) version 3.10.0-327.22.2.el7.x86_64
- Oracle Database 12c version 12.1.0.2.0 with Automatic Storage Management (ASM)
- Flexible IO (fio) Tester version 2.12-12. This optional open source software was used to verify performance of storage devices prior to creation of any Oracle databases.
- HammerDB version 2.20. This optional open source software was used to generate schemas and test data for the HammerDB-OLTP, HammerDB Power User Test and HammerDB Throughput Test benchmarks.
- VSL software, version 4.2.5. This is the driver for the Fusion ioMemory SX350 storage card.

To accommodate six Fusion ioMemory SX350 storage cards in the x3650 server, two PCIe risers were installed in the server. Each riser supports three full-height x8 PCIe cards. RAID controllers were not used in the server to control the storage cards.

Each storage card was presented to the operating system as an individual block storage device, so from a Linux perspective there were six storage cards named /dev/fio[a-f]. Each storage card was partitioned per Oracle ASM best practices, and permissions were established by a udev rule file as detailed below.

In this test effort, HammerDB was one of the benchmarks used alongside fio and Calibrate IO. HammerDB-OLTP is derived from the TPC-C benchmark and as such is not comparable to published TPC-C results. HammerDB Power User Test and HammerDB Throughput Test are derived from the TPC-H benchmark and as such are not comparable to published TPC-H results.

Understanding storage capacity

Each of the six Fusion ioMemory SX350 storage cards offers 6,400 GB of raw storage. The total storage capacity across all six storage cards is 38,400 GB.

GB vs GiB: A gigabyte (GB) here equals 1 billion bytes. Some applications, such as Oracle, use gibibytes1 (GiB, or 1,073,741,824 bytes). From this perspective, the capacity of each storage card would be 5,960 GiB, and the total capacity across all six storage cards would be 35,762 GiB. The difference in reported capacities is roughly 7.3% between base-ten GB and base-two GiB.

Using Oracle Advanced Compression could triple the amount of usable capacity to well over 100 TiB (tibibytes). Compression was not used in this solution, but the benefits are worth noting. In addition to storing more data in the same physical footprint, the data remains compressed when fetched into memory and therefore reduces memory requirements. Oracle Advanced Compression can also be used to reduce bandwidth requirements on network transmissions, as well as reducing the footprint of Data Pump Export files and Oracle Recovery Manager (RMAN) backups.

This solution evenly divided the available storage between two Oracle ASM disk groups named DATA and RECO. The size of each disk group was 17.4 TiB (or 18,310,542 MiB as reported by the asmcmd utility). All database files were stored on DATA, and all recovery files including Data Pump Export files were stored on RECO.

ASM Normal Redundancy was applied to each disk group, which is software-based RAID 10 protection. This reduced available storage by one half, so each disk group's usable space was 8.7 TiB. ASM Normal Redundancy also generates more I/O, since all writes are doubled. This level of protection was not required. Often, no redundancy is used for recovery files or for entire reporting systems, as the data is just a copy from the operational data store. This solution, however, enforced maximum protection by including ASM Normal Redundancy on all disk groups.

The HammerDB-OLTP schema used a scale of 10,000. This equates to roughly 5 billion rows consuming 1 TiB of space. However, the schema build required significantly more space, particularly while building indexes. The tablespace was sized at 2 TiB to accommodate the scratch space requirement.

The tablespace was not shrunk, because during testing there must be free space for row inserts. Each test inserts a considerable amount of new data, which eventually increased the size of the tablespace to 5TiB. The TEMP and UNDOTBS1 tablespaces were 100 GiB and 300 GiB, respectively.

The schema for the HammerDB Power User and HammerDB Throughput tests used a scale of 3,000. This equates to roughly 26 billion rows consuming 5 TiB of space. This scale corresponds to the available storage capacity. A larger schema was not loaded because it could double in size during testing due to the refresh function inserting new data.

Oracle Database systems generally have more storage than called for by a single requirement (such as performance). Capacity, redundancy, and other factors must also be addressed. For example, during the indexing portion of the HammerDB Power User and HammerDB Throughput schema build, simultaneous I/O rates of 2150 MB/s reads and 3050 MB/s writes occurred. The workload could have been handled by just two storage cards, but the other storage cards were needed for capacity and redundancy.

Building the system

Oracle Linux 7.2 with UEK was installed using the Server With GUI option. The packages and kernel were then updated by running yum update as root. Packages, users, groups, and system settings were implemented based on the Oracle online documentation. The kernel boot loader was configured for low-latency applications.

HugePages was configured for 50% of memory during OLTP testing, which ensures that the Oracle SGA remains pinned in memory and cannot be swapped out to disk. The SGA size is often smaller in DSS or analytics workloads, with a corresponding decrease in HugePages. Full table scans typically skip the SGA completely and require vast amounts of PGA memory. HugePages was shrunk to 26 GB during DSS testing, to maximize the Oracle PGA space.

All storage cards were formatted with a 512 byte sector size and partitioned with the Linux `parted` utility into two equal partitions of 3.2 TB. The first partition of each storage card was used to create the ASM diskgroup DATA, and the second partition on each storage card was used to create the ASM diskgroup RECO. The partitioning commands are shown in Figure 3.

```
for file in /dev/fio?
do
 parted -a optimal $file mklabel gpt mkpart primary 1 50% mkpart primary 50%
100%
 parted $file print
done
```

Figure 3 Partitioning commands

Note: This solution uses Oracle ASM 12c, which supports devices up to 32 PB each. The older Oracle ASM limits storage targets to 2 TB, in which case the parted utility can be used to divide a large storage device into many partitions each equal to or small than the 2 TB limit.

A very simple udev rules file was created to set ownership and permissions on all Fusion ioMemory SX350 storage cards. udev rules are automatically applied on system boot and can be modified and reapplied on a live system without interrupting applications like Oracle. The rules file is shown in Figure 4.

```
cat /etc/udev/rules.d/99-oracle.rules
KERNEL=="fio[a-z][1-9]", OWNER="grid", GROUP="asmadmin", MODE="0660"
```

Figure 4 udev rules

Note: ASMLib and ASM Filter Driver were not used for this paper, but they could replace the above udev rule if desired.

Oracle ASM was configured with two disk groups, DATA and RECO, for storing all database and recovery files. Redo logs were multiplexed across both diskgroups. Both diskgroups were created with default values: 512 byte sectors, 1 MB allocation unit size, and 12.1.0 compatibility. Both diskgroups used ASM Normal Redundancy, which is akin to RAID 10 data protection. Of course, the Oracle user is free to adjust ASM to meet individual requirements. This paper shows how excellent performance can be achieved without unusual settings.

Two Oracle databases were created: OLTP and DSS. Both used the AL32UTF8 character set, the Bigfile tablespace format, and five groups of multiplexed redo logs with 5GB member files (one member of each log group was created on each of the two ASM diskgroups). Other database settings used default values except as noted below.

- The OLTP database had an 8 KB block size and hosted the TPCC schema with a scale of 10,000 (roughly 5 billion rows of data plus the corresponding indexes). It’s important to keep the initial data set at a reasonable size, because new data created during testing will triple the schema size.
- The DSS database had a 32KB block size and hosted the TPCH schema with a scale of 3,000 (roughly 4 TB of data and indexes, plus 1 TB of temp and undo space). The data and index space can triple during testing, so it is important to leave plenty of free space.

Tip: HammerDB loads data in units called *scale*, but this term has a different meaning in each benchmark and does not directly determine storage requirements.

HammerDB-OLTP's scale is the number or rows in the WAREHOUSE table, and the other tables are populated with many rows per warehouse. (For example, the largest table, ORDER_LINE, has over 300,000 rows for each WAREHOUSE row.) The HammerDB Power User Test and HammerDB Throughput Test scale is the initial amount of data capacity in units of 1GiB per scale, but this does not account for indexes or new data loaded during testing. Common tablespaces such as TEMP and UNDO are also not included in scale estimates, and they can grow quite large during testing.

Each database ran separately, as is typically the case with OLTP and DSS databases, due to their different memory profiles. Multiple databases with similar profiles certainly could have been run simultaneously on this system, because it had ample resources. Users who need to host multiple databases should also consider Oracle Multitenant, which allows up to 252 databases to be run concurrently on a single system, all sharing one set of resources.

To illustrate Oracle Multitenant, imagine a bunch of grapes where the container database is the vine, and each application database is a grape. Oracle Multitenant maintains isolation of data and user accounts within each application database while consolidating some activities, such as backups and SGA management, at the container database level. Migrations and upgrades are simplified because databases can be unplugged from one container and plugged into another container. Each application database has its own DBA, while the container database has a super-DBA for global tasks.

Design choices

The choices of file system and volume manager are up to the user, as are the choices of partitioning and layout of the storage devices. Oracle offers many options to enhance their Oracle Enterprise Edition database, including compression, partitioning, and encryption.

This solution uses Oracle ASM for its balance of performance and ease of management. The storage cards could alternatively be mounted by a standard Linux file system such as EXT4 or XFS, with RAID provided by standard Linux utilities such as mdadm.

The Fusion ioMemory SX350 storage cards provide 6.4 TB of 100% NAND flash storage capacity with a performance profile that exceeds 170,000 database IOPS, 2800 MB/s sequential reads, and 2200 MB/s sequential writes. This combination of performance and capacity means a single storage card can easily accommodate multiple Oracle databases.

The Fusion ioMemory SX350 storage cards are often deployed in mirrored pairs. The x3650 M5 server supports up to six of these storage cards for very large database deployments and consolidation of many Oracle databases onto a single server.

Each storage card was split into two equal partitions:

- ASM disk group DATA partition
- ASM disk group RECO partition

This layout stripes the data evenly across all storage cards and maximizes bandwidth on both reads and writes in the DSS workload test. OLTP workloads, on the other hand, benefit more from IOPS and low latency than from bandwidth. OLTP testing showed the I/O performance requirements could be serviced by just one or two storage cards. There was no benefit to striping the data across all of the storage cards. So another way to lay out the storage, where the capacity is evenly divided between DATA and RECO would be to simply give the first three storage cards entirely to DATA and the next three storage cards entirely to RECO.

Oracle compression and partitioning were not used in this solution, but they could be implemented to enhance the system. Compression allows more rows to be stored per database block and processed within each memory page. Compression can also improve performance by allowing more rows to be transferred per I/O operation. For example, transferring an 8 KB block takes the same amount of time, regardless of the number of rows it contains.

No effort was made to isolate I/O profiles to specific storage devices. For example, no storage cards were dedicated for the Oracle redo log or archive log files. Instead, Oracle's best practice of Stripe And Mirror Everything (SAME) was used. This isn't possible with legacy storage, or even many of today's SSD sitting behind legacy storage protocols. It is possible with the Fusion ioMemory SX350 storage cards featured in this paper, because they were engineered to sustain heavy writes in concurrent user workloads.

Extending the architecture with Oracle Data Guard

The selected server and components are highly available. However, down time happens, such as when applying patches or performing various other maintenance activities. Oracle Data Guard can be used to minimize the impact of planned and unplanned down time by creating a copy of the database that is automatically synchronized with the original database.

Should the need arise to take the primary database offline, the roles of the two databases may be switched. Users will then connect to the database copy, and all work will be queued for later synchronization to the original database once it has been brought online. At that point, users can continue using the database copy (now considered the primary) or another switchover can be done, bringing users back to the original database.

Oracle refers to the original database as Primary, and refers to all copies as Standby. When a planned switchover or unplanned failover occurs, the role of Primary is transferred to the first available Standby, and that Standby database's role is transferred to the old Primary.

Oracle Data Guard is included with Oracle Database Enterprise Edition. Oracle Data Guard's capabilities can be extended through an additional license for Oracle Active Data Guard.

Oracle Active Data Guard allows users to connect with copies of the database that are actively receiving, as well as with the Primary. They can also perform read-only operations against these databases. This offers several benefits, including:

- Long running reports can be offloaded from the Primary database to any number of standby databases, thus freeing up resources on the Primary to service DML requests. (The read-only Oracle Active Data Guard database can be positioned anywhere in the world, bringing the data closer to users.)
- Oracle Recovery Manager (RMAN) backups can also be offloaded from the Primary to a Standby to avoid contention with user operations. Reader farms can be created in order to spread Internet scale user workloads across large numbers of read-only Oracle Active Data Guard databases.

To illustrate the benefits of Oracle Active Data Guard, consider a large data warehouse or analytics database. The underlying storage of any system has a maximum throughput. Let's assume that the storage layer of some system becomes saturated by X number of reports running concurrently. Each copy of the database with its own storage layer can run an equal number of reports concurrently. With one primary and one standby database one can run 2X reports, and with another standby one can run 3X reports, and so on. Thus, throughput scales linearly with the number of database copies.

Test procedures

The Fusion ioMemory SX350 storage cards were tested with two common Oracle Database workloads: Online Transaction Processing (OLTP) and Decision Support System (DSS). The Fusion ioMemory SX350 storage cards are well-known for low latency and high IOPS, which are very important to OLTP users. The high capacity and high I/O bandwidth of these storage cards supports database workloads such as reporting, data warehousing, and analytics characterized by long sequential reads and writes.

The test process is summarized below, and many of the activities are discussed elsewhere in this paper:

1. Install and configure storage cards.
2. Measure storage performance using fio (single storage card, then all storage cards together).
3. Create ASM disk groups DATA and RECO.
4. Create the OLTP database with an 8KB block size.
5. Measure database storage performance using Calibrate IO.
6. Load the “TPCC” OLTP schema using HammerDB.
7. Conduct OLTP performance testing using HammerDB.
8. Drop the OLTP database.
9. Create database DSS with 32K block size.
10. Measure performance using Calibrate IO.
11. Load the “TPCH” DSS schema using HammerDB.
12. Conduct DSS performance testing using HammerDB.

The HammerDB-OLTP benchmark was run for 90 minutes with 55 virtual users (two times physical core count) and again for 90 minutes with 150 users. The key and thinking time options were disabled in HammerDB, thus allowing the system to be flooded with transactions, and in effect simulating hundreds of thousands of users being funneled through connection pools.

The HammerDB Power User and HammerDB Throughput benchmarks do not have a fixed time. These benchmarks run until all users have completed the full set of predefined queries. The goal was to stress the I/O subsystem with long sequential reads and show how the high bandwidth of the Fusion ioMemory SX350 storage cards could expedite reporting and analytic workloads.

Test results

The sections below describe the various tests and their results. As noted earlier, three basic tools were used to measure performance: Flexible IO Tester (fio), Oracle Calibrate IO, and HammerDB.

fio test results

fio was used to measure performance of the storage cards (not databases) prior to mounting a file system (e.g., ASM) or creating any databases, to verify that the storage and system are working together correctly. Tests were run multiple times on each storage card separately to verify operational status. The tests were repeated using all storage cards in aggregate, again running the test multiple times. The fio test results were averaged for reporting. The table below shows the average results of testing all Fusion ioMemory SX350 storage cards with various fio tests.

Table 2 Averaged fio test results

	IOPS Average	Bandwidth Average
8KB Random Read	1,032,200	Not applicable
8KB Random Writes	1,207,500	Not applicable
8KB Mixed 60/40	726,174	Not applicable
1MB Sequential Reads	Not applicable	16,209
1MB Sequential Writes	Not applicable	13,992

Calibrate IO test results

Oracle Calibrate IO is a stored procedure included in all Oracle 11gR2 and higher databases. It can measure the performance of the portion of the storage subsystem allocated to application tablespaces. Storage dedicated to logs, backups, and other spaces was not utilized in the test. Calibrate IO is a read-only test that can be run at any time without concern of harming the database, but it is resource-intensive and should not be run on production systems during core business hours.

Prior to running Calibrate IO, a 1 TB tablespace was created in the ASM diskgroup DATA. Input parameters included the number of spindles in the storage layer, and the maximum tolerable latency was measured in milliseconds. Calibrate IO requires the latency input parameter be 10 ms or higher, so the minimum value of 10 was used. The “number of spindles” input parameter was set to 8X the number of storage cards. The results are displayed in Table 3, along with a brief explanation of each metric.

Table 3 Averaged Calibrate IO test results

Metric	Average Results 8K Block	Average Results 32K Block	Description
Random Read IOPS	1,052,935	301,783	This represents the highest level of database block reads sustained without violating the input parameter for latency. If a sub-test incurs latency higher than the latency input parameter allows, that sub-test is discarded. Of the sub-tests that are not discarded Oracle reports the one with the highest IOPS. The same sub-test is used for reporting latency.
Latency (Milliseconds)	0	0	This represents the latency incurred during the period for which max IOPS were reported. The value is reported only in whole milliseconds, so values from below 0.5 ms are rounded down to 0.

Metric	Average Results 8K Block	Average Results 32K Block	Description
Read-Only Bandwidth	16,386	16,371	This represents the highest level of throughput the database sustained for some threshold amount of time (Oracle does not disclose this threshold). All reads are 1MB sequential. This is not the absolute peak or an average; it is the highest sustained bandwidth. In this case the result was equivalent to 22 ports of 8Gb Fibre Channel[a] or 9 ports of 16Gb FC.

[a] Assumes 8 Gb Fibre Channel can sustain 750MB/s, and 16Gb Fibre Channel can sustain 1.8GB/s. The performance increase from 8Gb to 16Gb Fibre Channel is non-linear due to their different schemes for bit encoding.

Interestingly, the 32 KB block size delivered 14-15% more data per second in OLTP conditions compared to the default 8 KB block size, indicated by the IOPS results. Each I/O using a 32K block size delivered at least 4X the data of a typical 8 KB block (slightly more, because there are only one-fourth as many block headers).

Multiplying the IOPS number times the block size shows that the 32 KB block size database delivered 9.89 GB/s, compared to 8.63 GB/s for the 8K block size database. Both databases delivered I/O with sub-millisecond latency. **Note:** Users should keep these points in mind when choosing a block size for OLTP databases. Higher block sizes are often avoided due to the higher latency on spinning disk storage solutions. This is *not* an issue on modern flash storage solutions.

About the HammerDB test results

The HammerDB-OLTP, HammerDB Power User and HammerDB Throughput workload results are noted in the next sections. The system was configured for real-world use cases. For example, standard Oracle Database features such as ASH and ADDM that consume resources were *not* disabled, and unprotected storage was not used: all storage was configured with ASM Normal Redundancy.

HammerDB is an open source product that can be used to compare performance of a given database on various platforms. For example, one might use HammerDB to determine a level of performance using one set of processors, then swap out the processors for new ones and rerun the test. It can also be used to assess the impact of certain types of database changes, such as increasing memory or resizing the log files.

Keep in mind, there are many variables that make it difficult to compare one user's HammerDB results with another user's results. For example, users might obtain different results due to adjustments they have made to the operating system, network, database, or the HammerDB configuration. For that reason, customizations in this solution were kept to a minimum and are noted in this document so the tests can be easily repeated for independent comparisons.

The HammerDB tests offers workloads based on popular benchmarks: TPC-C and TPC-H. The HammerDB tests are not official TPC benchmarks, but they use a compatible schema and logic:

- The HammerDB-OLTP workload is characterized by many users trying to place as many new orders into the system as possible per unit of time, while also performing various other tasks. This workload stresses the CPU and database internals, and it is impacted by the latency on storage devices. The HammerDB-OLTP workload performs single block random reads and writes.

- The HammerDB Power User and HammerDB Throughput workloads are characterized by a few users with parallel processes searching vast amounts of data, which stresses the I/O sub-system in particular, and is greatly impacted by the bandwidth in the storage layer. The HammerDB Power User and HammerDB Throughput workloads perform long sequential reads and bulk inserts.

HammerDB-OLTP test results

The system exceeded 2 million HammerDB transactions per minute in all runs of the HammerDB-OLTP workload, with CPU utilization in the range of 75-85%. The storage layer maintained submillisecond latency on random reads and redo log writes, critical to Oracle Database performance.

HammerDB provides two metrics:

- TPM, which is the average number of transactions completed per minute,
- NOPM, which is the average number of new orders processed per minute.

TPM can be easily compared to other Oracle databases, while NOPM allows for comparison to non-Oracle databases. TPM can be quickly calculated as the total number of commits and rollbacks during the test period divided by the number of minutes in the test period. NOPM is the number of executions of the New Order procedure during the test period divided by the number of minutes in the test period. Both the TPM and NOPM metrics are provided in the HammerDB log file, along with AWR snapshot numbers.

The Oracle AWR report provides database, system, and storage performance metrics. At the end of each HammerDB-OLTP test, the HammerDB log file identifies the two AWR snapshots that were created before and after the test. The DBA can use these two snapshot numbers to generate an AWR report. The report's key metrics include the following, which are shown in Table 4 on page 15 with actual test values.

Table 4 Key metrics and test values

Report metric	Test values
Host CPU %idle	22.0%
db file sequential read	0.22 ms
db file scattered read	1.22 ms
db file parallel write	0.28 ms
log file parallel write	0.12 ms

Using the transaction monitor: HammerDB offers an on-screen transaction monitor. While it may be interesting to watch, the transaction count is misleading and should be ignored. For example, this graph's reliability was tested using a commodity SSD:

- With refresh interval of 1 second, the HammerDB monitor reported 4.1 million TPM
- With refresh interval of 10 seconds, the HammerDB monitor reported 2.2 million TPM
- With refresh interval of 45 seconds, the HammerDB monitor reported 1.4 million TPM

The test's actual performance was 1.3 million TPM.

HammerDB Power User test results

A harsh testing environment was purposefully created where the amount of data was 16 times the amount of available system memory. This was done to show the benefits of flash storage without skew from data caching.

The Power User Test has a single user running all 22 predefined queries, with a high degree of parallelism. The DOP was set to 56, which is twice the number of physical CPU cores in the server. The Power User Test ran three times to account for variations caused by randomly selected predicates. Occasionally, the random value passed to the predicate would result in significantly more data being fetched, while other values would fetch less data.

The result was high CPU utilization and high storage bandwidth utilization with 15 trillion 32KB-block reads at speeds up to 16 GB/s. CPU utilization throughout the test ranged from 45% to 100%, with the average utilization being quite high. (Recall that the degree of parallelism was set to twice the number of CPU cores.) The storage layer returned data at rates up to 16 GB/s, with 15 GB/s typically observed on all queries. Total block reads from the data tablespace TPCHTAB was 15 trillion 32KB blocks during the three iterations of the Power User Test.

Performance can be improved by adjusting the ratio of RAM to storage data. The test used a ratio of 1:16, which is quite low. Additional RAM would allow data to be processed more fully in memory, and less in swap or temp space. Tablespace TEMP grew to 971 GB during testing.

Another way to improve performance of the DSS workload is to compress the data. Because each block would hold 3-4 times as much data, each I/O should return 3-4 times as much data, and the queries would complete that much sooner. The reader should consider using Oracle Advanced Compression.

HammerDB Throughput test results

The Throughput Test has multiple concurrent users running massive queries with parallelism. Testing ran with 8 users, which is typical for databases of this scale.

The Throughput Test used the same set of predefined queries as were run in the Power User Test. The order in which each user executed the 22 pre-defined queries is shown in Table 5 and Table 6.

Table 5 HammerDB Throughput query execution ordering (part 1)

Query	1	2	3	4	5	6	7	8	9	10	11
User 1	21	3	18	5	11	7	6	20	17	12	16
User 2	6	17	14	16	19	10	9	2	15	8	5
User 3	8	5	4	6	17	7	1	18	22	14	9
User 4	5	21	14	19	15	17	12	6	4	9	8
User 5	21	15	4	6	7	16	19	18	14	22	11
User 6	10	3	15	13	6	8	9	7	4	11	22
User 7	18	8	20	21	2	4	22	17	1	11	9
User 8	19	1	15	17	5	8	9	12	14	7	4

Table 6 HammerDB Throughput query execution ordering (part 2)

Query	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
User 1	15	13	10	2	8	14	19	9	22	1	4
User 2	22	12	7	13	18	1	4	20	3	11	21
User 3	10	15	11	20	2	21	19	13	16	12	3
User 4	16	11	2	10	18	1	13	7	22	3	20
User 5	13	3	1	2	5	8	20	12	17	10	9
User 6	18	12	1	5	16	2	14	19	20	17	21
User 7	19	3	13	5	7	10	16	6	14	15	12
User 8	3	20	16	6	22	10	13	2	21	18	11

By design, several users ran the same query at the same time, causing contention for resources. Also, because some queries take more time than others, users can “catch up” to each other, resulting in even more users running the same query concurrently. The system under test must be capable of maintaining good performance under these conditions.

The test system maintained excellent performance on the Throughput Test. CPU utilization ranged from 60 to 95%, with an average utilization of 80%. I/O bandwidth was maintained in the range of 10 to 16 GB/s.

References

See the following links for more information:

- Lenovo Press Product Guide for the Lenovo System x3650 M5:

  https://lenovopress.com/lp0068-lenovo-system-x3650-m5-e5-2600-v4
- Lenovo Press Product Guide for the SanDisk Fusion ioMemory SX350 Application Accelerator:

  https://lenovopress.com/lp0058-io3-enterprise-mainstream-flash-adapters
- Oracle Database 12c:

  https://www.oracle.com/database
- HammerDB:

  https://sourceforge.net/projects/hammerora/
- gibibytes1:

  https://en.wikipedia.org/wiki/Gibibyte

Conclusions

Running an Oracle Database 12c, the x3650 M5 configured with SanDisk Fusion ioMemory SX350 provides an affordable, small-footprint enterprise solution that delivers both high reliability and performance. To reduce costs and achieve a rapid return on investment, enterprises can run large performance-sensitive Oracle databases with confidence, or consolidate databases from many servers.

The system exceeded 2 million HammerDB transactions per minute in all runs of the HammerDB-OLTP workload, with CPU utilization in the range of 75-85%. The storage layer maintained sub-millisecond latency on random reads and redo log writes, which is critical to Oracle Database performance.

The result of the HammerDB Power User Test was high CPU utilization and high storage bandwidth utilization with 15 trillion 32 KB-block reads at speeds up to 16 GB/s. CPU utilization throughout the test ranged from 45% to 100%, with the average utilization being quite high (recall that the degree of parallelism was set to twice the number of CPU cores). The storage layer returned data at rates up to 16 GB/s, with 15 GB/s typically observed on all queries. Total block reads from the data tablespace TPCHTAB was 15 trillion 32 KB blocks during the three iterations of the Power User Test.

Under the HammerDB Throughput Test the system maintained excellent performance. CPU utilization ranged from 60 to 95%, with an average utilization of 80%. I/O bandwidth was maintained in the range of 10 to 16GB/s.

This system is well-suited to large Oracle Database 12c deployments running OLTP or DSS workloads. The solution also facilitates server and database consolidation, which reduces operating expenses (OPEX) and provides a more rapid return on investment.

As discussed in this paper, server consolidation is a great way to reduce operating costs by either running multiple, physically separate databases per server or using Oracle Multitenant introduced in Oracle Database 12c. The Oracle Database 12c user can achieve maximum efficiency and savings by deploying Oracle Multitenant on flash-powered servers like the x3650 M5 with Fusion ioMemory SX350 storage.

Change history

February 2017:

- Grammar and readability improvements

Authors

This paper was produced by the following team of specialists:

Mark Johnson is an Oracle 12c Certified Professional DBA with a Master’s Degree in Information Management and over 20 years’ professional experience in Oracle database technology. Mark works in the SanDisk Data Propulsion Labs where he is responsible for Oracle performance studies and solution designs.

Prasad Venkatachar is the Senior Solutions Product Manager for Databases and Big Data with Lenovo where he focuses on building industry leading Lenovo solutions. He has extensive experience in IT Services, Presales and Marketing. His technical expertise include Oracle/DB2 and Big Data Hadoop and NoSQL. He is certified on Oracle database releases from 8i to 12c and also certified on the IBM DB2, ITIL V3, and VMware Virtualization products.

Ron Kunkel has over 25 years of data center experience working with IP networks, telecommunication products/applications and data center electrical power systems. In these fields, he has operated in a diverse range of roles which include: Product Manager, Product and Solutions Marketing, Business Development, Solution Sales Engineering and Design Engineer.

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

- David Watts, Lenovo Press
- Mark T. Chapman, Lenovo

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.

This document was created or updated on February 3, 2017.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp0586

Trademarks