StorageReview

NetApp
Memory Accelerated Data

Solution Brief

December 2018

Contents

Executive Summary

NetApp Memory Accelerated Data (MAX Data) is a server-side technology designed to accelerate application performance in an entirely non-disruptive way. With MAX Data, the enterprise has the opportunity to deploy the latest, and fastest, storage technologies without having to rewrite applications or institute complicated architectures that are difficult to deploy and support. Furthermore, MAX Data extends the data protection, persistence, and tiering capabilities of ONTAP between the server and the traditional NetApp AFF storage array.

NetApp brings MAX Data to fruition by leveraging persistent memory (PMEM) or DRAM within the host-application server. While the DRAM and NVDIMM solutions are available today, NetApp is also leading the pack as far as making PMEM a reality in the enterprise space. The exciting value of PMEM is that it enables performance that's faster than local storage and less expensive than DRAM. This means MAX Data is not only cost effective because it uses fewer application servers and DRAM, but it also delivers immense improvements to application responsiveness.

Benefits of MAX Data

- Server side software accelerates performance with no changes needed to the application
- ONTAP data services across MAX Data server and AFF storage array
- Latency improvement of up to 13.5X as tested in single server configuration
- Fewer application servers and DRAM needed
- Fast application data recovery with MAX Recovery

While most applications can benefit from MAX Data, the emphasis is clearly on the hungriest use cases like real time analytics, trading platforms, artificial intelligence, in-memory databases and data warehousing. In all of these cases, there's an element of time sensitivity and criticality that is crucial to business operations. By applying MAX Data to the applications common in these use cases, organizations are able to shrink processing time. Shorter processing times means business decisions can be made more quickly, which is a tremendous competitive advantage.

In this solution brief, we detail the architecture and benefits of MAX data, bringing clarity to NetApp's use of PMEM and capabilities within the current shipping version 1.1. We also discuss the real-world benefits applications can see from MAX Data by leveraging the SLOB2 tool to illustrate Oracle Database performance capabilities. Data is compared between a host application server with and without MAX Data in place, illustrating the real-world benefit of MAX Data.

What is MAX Data?

It is crucial for mission critical systems such as in-memory databases, artificial intelligence (AI), machine learning (ML), and bespoken high-frequency trading and fraud detection applications to be able to access and process data as quickly as possible. Because every millisecond of latency within these types of applications can make a profound difference in the success or failure of an operation, companies that use them are constantly looking for ways to improve their response times.

New hardware breakthroughs have emerged in response to these requirements for ultra-low latency applications. One software technology poised to take advantage of these is NetApp MAX Data (short for Memory Accelerated Data), as MAX Data enables the use of persistent memory (PMEM) on a server to deliver consistent ultra-low latency for persistent data storage. It does so by using storage class memory (SCM) to create a high-performance auto-tiering file system. SCM is a new class of memory that is just now starting to come into common use. The performance characteristics approach that of dynamic random-access memory (DRAM), but with persistence. This means that the data stored on SCM, unlike DRAM, will still be present after a system reboot.

Given the high stakes nature of an operation, trusting mission critical applications to a new and unproven technology is a no-go for most enterprises. However, this should be of minimal concern to enterprises looking to implement MAX Data in their datacenter. Even though it was only just officially announced in 2018, its underlying technology has been under development and refinement since 2013, as NetApp built upon the technology and intellectual property (IP) of Plexistor, an interesting company that was slightly ahead of its time. In fact, the hardware to enable the technology to reach its full potential was just starting to come on the market when NetApp shrewdly acquired the company in 2017.

By using Plexistor technology as a starting point, NetApp developed the heart of MAX Data, MAX FS—a high-performance auto-tiering Linux file system. The file system consists of a primary tier residing on the server and secondary storage residing on a NetApp Array. Applications do not need to be rewritten to take advantage of MAX FS as it is POSIX-compliant; virtually all applications and data that currently reside on a Linux file system will be able to be used on MAX FS without modification. MAX Data not only supplies the file system, but also a custom driver to maximize the potential of MAX FS. This driver strips away the unneeded layers that traditional Linux file systems rely upon. As a result, the performance benefits of SCM to drive down latency are yielded. In fact, NetApp touts single-digit millisecond latency of applications using MAX Data.

NetApp presented a chart (Figure 1) during a session at NetApp Insight 2018 which compares a MAX FS file system (identified in the chart as M1FS) to a traditional Linux file system (XFS) when running MongoDB. This chart shows that although MAX Data is more than 3 times faster under a light load, it really shines under a heavy load when it can complete 11 times more documents per query than XFS. Although the specifics of the testing are outside this paper's scope, it does demonstrate that MAX Data has the potential to radically increase an application's performance.

What is MAX Data?

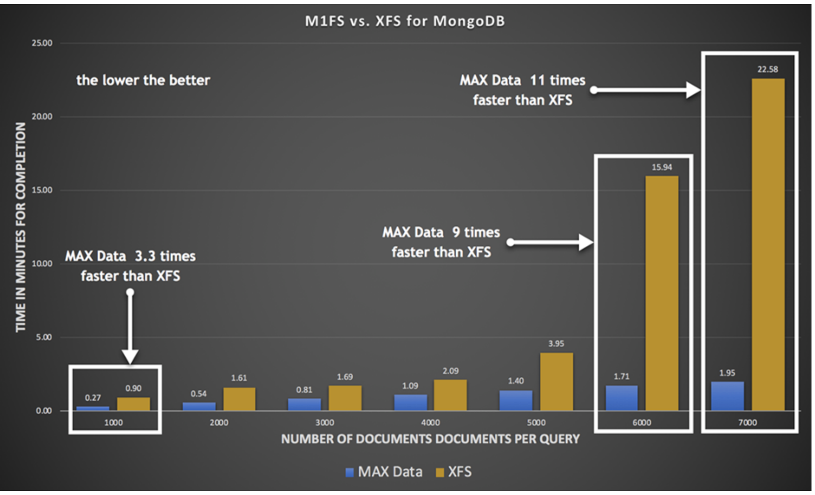
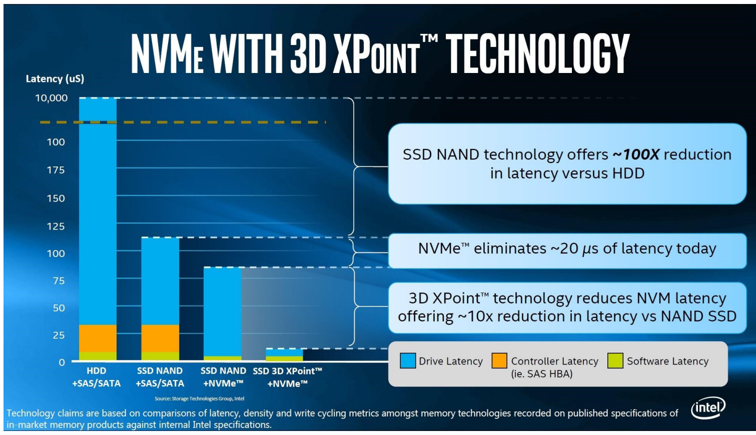


Figure 1

To briefly summarize, MAX Data is able to combine the capacity needed for persistent storage with the latency of real memory. MAX Data is a tiered file system where data resides in either the persistent memory tier or the storage tier. Sized properly, the application's most-accessed data resides in the memory tier, while the warm and cold data reside on the storage tier. Data that is unused on the persistent memory tier is moved to the storage tier and data that is needed from the storage tier is moved to the persistent memory tier, allowing hotter data to have more space.

The current version of MAX Data (1.1) uses NVDIMM or DRAM backing the MAX FS file system. As DRAM is non-persistent, its use cases with MAX Data are niche and rather limited, but NVDIMM (non-volatile DIMM) is persistent and is classified as SCM. However, NVDIMM is rather expensive, requires specialized motherboards, and is limited in its storage capacity. The real beauty in MAX Data is shown on NetApp's short-term roadmap; the company plans to use Intel Optane DC for the PMEM. 3D XPoint DIMM has similar performance characteristics of DRAM, but with persistence. And as a bonus, it comes at a lower cost, and is scheduled to have larger capacity than NVDIMM or DRAM. Intel is currently sampling 3D XPoint DIMMs and is shipping them to select customers, with general availability slated for 2019. Intel has released a chart (Figure 2) comparing 3D XPoint technology with traditional storage; although the chart is based on non-volatile memory express (NVMe), these numbers can be used to get an overall idea of how well 3D XPoint will perform with a PMEM solution such as MAX Data. The contrast in latency between traditional storage and 3D XPoint is remarkable. More specifically, the chart shows ~10 microseconds in latency for 3D XPoint, while SSD NAND SATA devices are closer to 100 microseconds, and even NVMe NAND devices are around 75 microseconds. The key takeaway here is that SCM can deliver data in microseconds, while traditional storage devices deliver data in orders of magnitude slower.



SSD NAND technology offers ***~100X*** reduction in latency versus HDD

NVMe™ eliminates ~20 μs of latency today

3D XPoint™ technology reduces NVM latency offering ~10x reduction in latency vs NAND SSD

Source: Storage Technologies Group, Intel

Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

Figure 2

What is MAX Data?

Optane is the Intel brand name of products that uses 3D XPoint technology. MAX Data is positioned to provide customers a way to seamlessly take advantage of these technologies without having to re-architect existing applications. MAX Data also will enable rapid adoption of persistent memory by new applications that can leverage the flexibility of a POSIX-compliant file system consumable at memory speeds.

Intel isn't the only supplier of SCM hardware; in fact, Micron, which co-developed 3D XPoint with Intel, brands its 3D XPoint products QuantX. Samsung's SCM product is marketed as Z-SSD, which uses a different technology than 3D XPoint, but still delivers ultra-low latency. Other, more exotic, SCM technologies are still in development, including nanotube RAM (NRAM) and resistive RAM (ReRAM). NetApp's roadmap for MAX Data indicates that Optane will be supported first. It isn't too far of a stretch though to imagine how NetApp could easily incorporate these other SCM devices into MAX Data in the future.

MAX Data is supported on a few different versions of Linux, and on various servers. You will also need ONTAP 9.5 and a NetApp array that has been certified to work with MAX Data. No exotic hardware or software other than a large amount of DRAM or a PMEM device that plugs into a server's memory bus is required to run MAX Data. A current list of the supported hardware and software for MAX Data is found using the NetApp interoperability Matrix Tool (IMT) at https://mysupport.netapp.com/matrix

MAX Data is Enterprise Ready

Performance of MAX Data is only part of the story. Enterprises need more than ultra-low latency for mission-critical data storage. They also need data resilience and data efficiency. NetApp has addressed this need by announcing that ONTAP 9.5 will enable MAX Data to be enterprise-ready. We can see in the NetApp supplied diagram (Figure 2) that MAX Data can be coupled with other NetApp technologies to meet an enterprise's requirements.

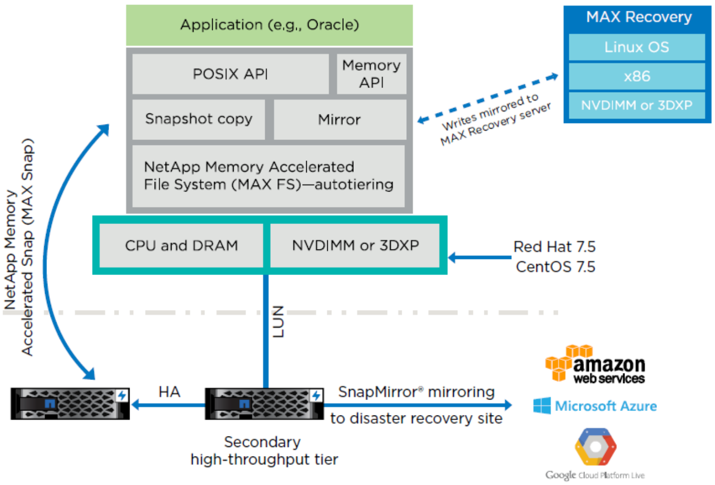

Application (e.g., Oracle)
POSIX API
Memory API
Snapshot copy
Mirror
NetApp Memory Accelerated File System (MAX FS)—autotiering
CPU and DRAM
NVDIMM or 3DXP
MAX Recovery
Linux OS
x86
NVDIMM or 3DXP
Writes mirrored to MAX Recovery server
Red Hat 7.5
CentOS 7.5
NetApp Memory Accelerated Snap (MAX Snap)
LUN
HA
SnapMirror® mirroring to disaster recovery site
Secondary high-throughput tier
amazon web services
Microsoft Azure
Google Cloud Platform Live

Figure 3

MAX Recovery mirrors the MAX Data file system to a MAX Recovery node via a private RDMA Ethernet connection to provide data reliance and last-transaction protection for databases. Each MAX Recovery node can support up to four MAX Data file systems.

What is MAX Data?

In the event of a failure of a MAX Data server, the data stored on the MAX Recovery node is mirrored back to the original MAX Data server or to a standby MAX Data server. Due to the performance characteristics of the MAX Data and the high-speed network requirements for MAX Recovery, the restoration of data using MAX Recovery takes a fraction of the time it takes using traditional data recovery methods.

Although MAX Data greatly expands the amount of data that can be stored in PMEM, some of data may eventually become cold and will need to be accessed less frequently. When the primary tier reaches 90% capacity, the least recently used data will be auto-tiered to a secondary tier of less expensive and less performant storage provided by a NetApp-based All Flash FAS (AFF) in order to free up space for new or more active data on the primary tier of storage. All writes will first go to the primary, local servers file system. NetApp specifies that the secondary storage tier is 25 times the size of the primary storage tier. The sizing of the primary tier is based on (in the case of a database) the working set of the database and the amount of data misses that can be tolerated. The sizing for other uses and applications will depend on the amount of data that needs to be accessed via an ultra-low latency storage system, and the amount of data that can tolerate a higher latency.

ONTAP data management's features include high availability (HA), cloning, and Snapshot copies. Also, ONTAP SVM-DR for automated disaster recovery (DR) can be used with MAX Data to protect the data that resides on it.

MAX Data has two different licensing tiers: Basic and Advanced. Regardless of the license level, MAX Data is licensed per server, per year. MAX Data Advanced has all the features of the basic tier, but also supports MAX Recovery.

MAX Data is an important technology in NetApp's portfolio. NetApp's future vision for MAX Data, outlined on their roadmap, includes collaboration with other integrated cloud services including ONTAP Fabric Pool Archive/Offsite, ONTAP NDAS Backup, NPS, and Cloud Volumes ONTAP. Of course, as is always the case with vendors, all future plans have the caveat that they are subject to change, but these plans do fit in nicely with NetApp's data fabric vision.

NetApp is in a unique position with MAX Data. They couple it with ONTAP data management capabilities to deliver high performance, ultra-low latency, resilient, and reliable storage for in-memory systems, such as databases, or other applications that require ultra-low latency and enterprise-level protection. In sum, NetApp MAX Data takes advantage of persistent memory, such as Intel Optane DC, to wrangle every bit of performance from your current enterprise applications.

Solution Configuration

To show the performance enhancements offered by NetApp MAX Data within a single server, an environment was built to show like-to-like results. The infrastructure consists of a pair of well-equipped x86 servers, 32Gb Fibre Channel fabric powered by a Brocade G620 FC switch and a NetApp AFF A300 HA pair with one logical interface (LIF) per port. Each node in the HA pair has a single aggregate and features 24 drive partitions using NetApp's Advanced Drive Partitioning (ADP). Both sets of LUNs are presented over a single initiator group to the standalone Oracle server through one Storage Virtual Machine (SVM).

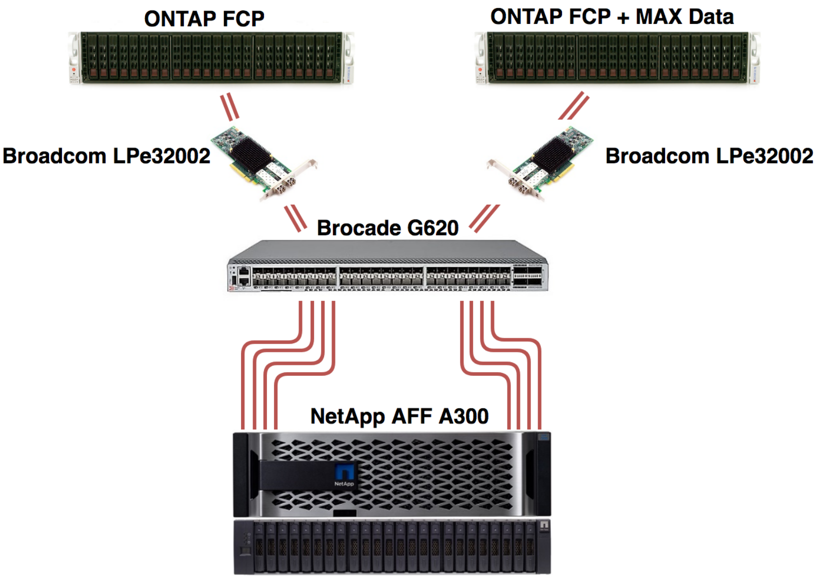


Figure 4

Servers

- x86 Server with X11DPi-N Motherboard
- 2 x Intel 6154 CPU (3.0GHz, 18 Core, 24.75MB Cache)
- 16 x 64GB DDR4-2666MHz ECC LRDIMM
- 480GB M.2 NVMe Boot SSD
- Broadcom LPe32002 32Gb Fibre Channel Host Adapter
- Mellanox ConnectX-5 dual-port 100GbE NIC MCX416A-
- CCAT
- Red Hat Enterprise Linux Server release 7.5 (Maipo)

Storage

- NetApp AFF A300 HA Pair
- 24 x 960GB SAS3 SSDs
- Dual Aggregates, one per controller, with 24 disk partitions each
- DP-RAID, Inline Compression and Deduplication Enabled
- 8 x 32Gb FC Links, 8 LIFS
- ONTAP 9.5

The Oracle server measuring ONTAP FCP performance has Oracle 18c and Grid 18c installed in standalone server configuration. From the storage array, 32 200GB FCP LUNs are presented, 16 per node, hosting the Oracle datafile storage. These nodes also presented 4 additional LUNs, 2 per node, used to host the redo log space for the database. Oracle Automatic Storage Management (ASM) is used to create disk groups for the Datafiles and Logs. ASM handles distributing data across the disks in the disk groups and provides optimal performance, balancing datafiles across the environment.

Solution Configuration

For the Oracle server measuring ONTAP FCP + MAX Data performance, Oracle 18c is installed in a standalone server configuration. MAX Data provisions the application, backed by four FCP LUNs to provide a storage tier to extend the mimic persistent memory tier using Application Aware Data Management (AppDM). MAX Data handles configuring igroups for the application. Separately, from the storage array, an additional four 400GB FCP LUNs are provisioned for Redo logs, two per node, to be managed by MAX Data in a single mount point to contain the datafiles. These form a volume group and logical volume using Logical Volume Manager (LVM2). Next, an XFS file system is formed on this new 1.6T logical volume. This hosts redo logs and control files for the MD database. A database is then provisioned with data files residing on the MAXFS and redo logs residing on this XFS file system.

On both servers, one SLOB schema is provisioned with SCALE=768G at 0 and 25% update levels at a range of thread counts to fully explore the IO profile of this storage configuration in test. This ensures that the working set size remains consistent across the entire test. A 1000GB SLOB tablespace is created to hold the SLOB schemas, which reside on the Datafiles diskgroup/LUNs.

After the tests complete, data is extracted from Oracle Automatic Workload Repository (AWR) reports at different load levels to get the read response time and total physical IOPS at each load level. From this data, the response time from ONTAP or from MAX Data at a given IOPS level is extracted. This provides a common comparison point that is relevant to any Oracle database.

In both test cases, the Oracle System Global Area (SGA) size is restricted to ensure that data is served by the underlying storage. This provides an accurate comparison of the capabilities of the storage tier instead of an exercise in Oracle-read caching performance.

Hardware Configuration

SLOB Overview

While there are a variety of tools that can be used to test Oracle databases, they all tend to fall short in one area or another. For instance, there are several transactional benchmarks that fail to properly assess Oracle's random physical I/O capabilities. The I/O benchmarks that can be used for Oracle are not database I/O and can give users a false sense of security. To address these issues, SLOB was created.

SLOB is not a database benchmark; it is a free tool providing Oracle I/O workload generation. SLOB is a middle ground between I/O generating benchmarks and full-function transactional benchmarks designed to the full Oracle database stack. SLOB is an SGA-intensive workload kit that creates Oracle SGA-buffered physical I/O without combining it with application contention. Since Oracle SGA physical I/O leverages cache miss, SLOB scales logical I/O allowing it to scale cache miss and have no issue when cache misses cross paths with cache hits.

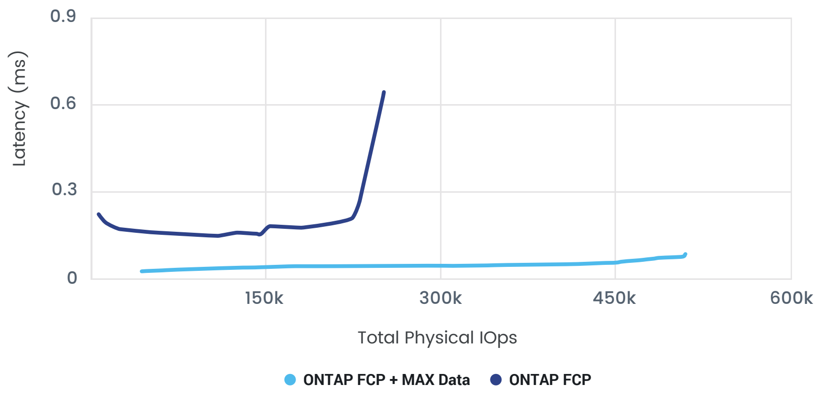
SLOB supports testing database performance or underlying storage performance. By manipulating settings such as SGA size, we are able to drive storage-centric testing through the Oracle stack. For example, a user can run the smallest db_cahche_size to run nothing but random single block reads. And, of course, users can drive up I/O as that is one of the main benefits of SLOB.

The SLOB toolkit is available for download at https://kevinclosson.net/slob/. The kit includes the following:

- README files and documentation. After extracting the SLOB tar archive, users can find the documentation under the "doc" directory in PDF form.
- A simple database creation kit. SLOB requires very little by way of database resources. The best approach to testing SLOB is to use the simple database creation kit under ~/misc/create_database_kit. The directory contains a README to assist the user. Use the simple database creation kit to create a small database because it uses Oracle Managed Files so users can simply point it to the ASM diskgroup or file system they want to test. The entire database will need no more than 10 gigabytes.
- An IPC semaphore based trigger kit. This kit does require permissions to create a semaphore set with a single semaphore. The README-FIRST file details what users need to do to have a functional trigger.
- The workload scripts. The setup script is aptly named setup.sh and to run the workload, users will use runit.sh. These scripts are covered in README-FIRST.

StorageReview

Performance Results


Oracle - 768GB Database, SLOB 100% Select

The advantages of running NetApp MAX Data become clear when you see the dramatic latency decrease between traditional ONTAP FCP performance from an AFF A300 with or without Max Data enabled on the host. In the first chart with a 100% Select workload on an Oracle 768GB database, MAX Data is able to drive higher throughput at a much lower latency than without. Peak throughput is nearly doubled, while latency across the full MAX Data run scales from 16 to 57 microseconds. This is in contrast to the underlying A300 storage measuring 216 to 634 microseconds by itself.

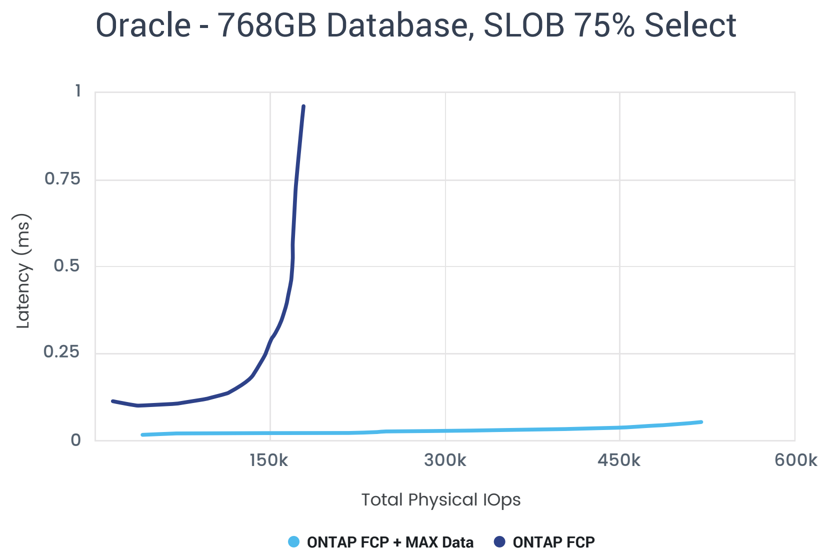
	Low Concurrency[1]	High Concurrency[2]
ONTAP	**216 µs**	**169 µs**
MAX Data	**16 µs**	**52 µs**
Improvement	**13.5x**	**3.25x**

[1] Low Concurrency = 1 SLOB Thread

[2] High Concurrency = 32 SLOB Threads

Looking at low and high concurrency in the 100% Select workload from SLOB, ONTAP FCP without MAX Data at a level of 1 thread measured 216 microseconds. In contrast ONTAP FCP with MAX Data at 1 thread measured just 16 microseconds or, in other words, a 13.5x improvement. At a high concurrency level of 32 threads, ONTAP FCP measured 169 microseconds, whereas with ONTAP FCP with MAX Data measured 52 microseconds, a 3.25x improvement.

Performance Results



Transitioning from a 100% Select to a 75% Select SLOB workload, the benefits of MAX Data continue to improve. With the same Oracle database size of 768GB, ONTAP FCP with MAX Data offered nearly triple the peak throughput with latency that scales from 17 microseconds to just 58 microseconds. Without MAX Data the same underlying A300 storage array scaled from 118 to 955 microseconds.

	Low Concurrency[1]	High Concurrency[2]
ONTAP	**118 µs**	**471 µs**
MAX Data	**17 µs**	**58 µs**
Improvement	**6.94x**	**8.12x**

[1] Low Concurrency = 1 SLOB Thread

[2] High Concurrency = 32 SLOB Threads

Looking at a low and high concurrency table for the 75% Select workload from SLOB, ONTAP FCP without MAX Data at a level of 1 thread measured 118 microseconds. With MAX Data, at the same 1 thread load level, that latency dropped to 17 microseconds, or a 6.94x improvement. At a high concurrency level of 32 threads, ONTAP FCP without MAX Data measured 471 microseconds. With MAX Data, at the 32 thread load level, latency measured 58 microseconds or a 8.12x improvement.

Conclusion

With MAX Data, NetApp continues to prove it's a leader in adopting emerging technologies to accelerate applications, thus enabling rapid adoption with enterprise performance and reliability. Beyond acceleration, MAX Data offers rich data management services, agile emerging technology compatibility and above all else, incredible performance benefits. Often new and innovative technologies require highly disruptive processes like re-architecting an application or datacenter infrastructure. With MAX Data this is not the case, as the testing required no application rewrites in order to see a huge improvement in application responsiveness.

Using MAX Data while running a SLOB 100% and 75% Select test on Oracle, we saw up to an order of magnitude improvement in low concurrency tests and up to an 8x improvement in high concurrency tests. Of course, this performance would be meaningless if it didn't have enterprise level management data services, which MAX Data does. MAX Data offers data protection and resiliency benefits ONTAP has offered in traditional storage arrays.

Currently MAX Data (v1.1) supports DRAM and NVDIMMs but the real game changer for this technology is Optane DC Persistent Memory, which is the obvious target for NetApp. When Intel's Optane DC PMEM becomes readily available, it will remove many of the cost and implementation barriers to this technology. As a result, MAX Data will be more cost effective to deploy and manage.

Enterprises that are looking to extract the maximum performance from their database or other latency sensitive applications should look at MAX Data. Performance may be the draw to MAX Data but when it is coupled with ONTAP to provide data protection and resiliency, MAX Data becomes an enterprise-ready application acceleration solution.

Appendix

Oracle SLOB Parameters

```
UPDATE_PCT=0
RUN_TIME=480
WORK_LOOP=0
SCALE=768G
WORK_UNIT=16
REDO_STRESS=LITE
DATABASE_STATISTICS_TYPE=awr
LOAD_PARALLEL_DEGREE=4
THREADS_PER_SCHEMA=1
DO_HOTSPOT=FALSE
HOTSPOT_MB=8
HOTSPOT_OFFSET_MB=16
HOTSPOT_FREQUENCY=3
HOT_SCHEMA_FREQUENCY=0
THINK_TM_FREQUENCY=0
THINK_TM_MIN=.1
THINK_TM_MAX=.5
NO_OS_PERF_DATA=1
DBA_PRIV_USER=”system”
SYSDBA_PASSWD=”oracle”
export UPDATE_PCT RUN_TIME WORK_LOOP SCALE WORK_UNIT
LOAD_PARALLEL_DEGREE REDO_STRESS
export DO_HOTSPOT HOTSPOT_MB HOTSPOT_OFFSET_MB NO_OS_PERF_DATA
export HOTSPOT_FREQUENCY HOT_SCHEMA_FREQUENCY
THINK_TM_FREQUENCY THINK_TM_MIN THINK_TM_MAX
```