Introducing Memory-Accelerated Data for FlexPod

Last Updated: October 1, 2019

Contents
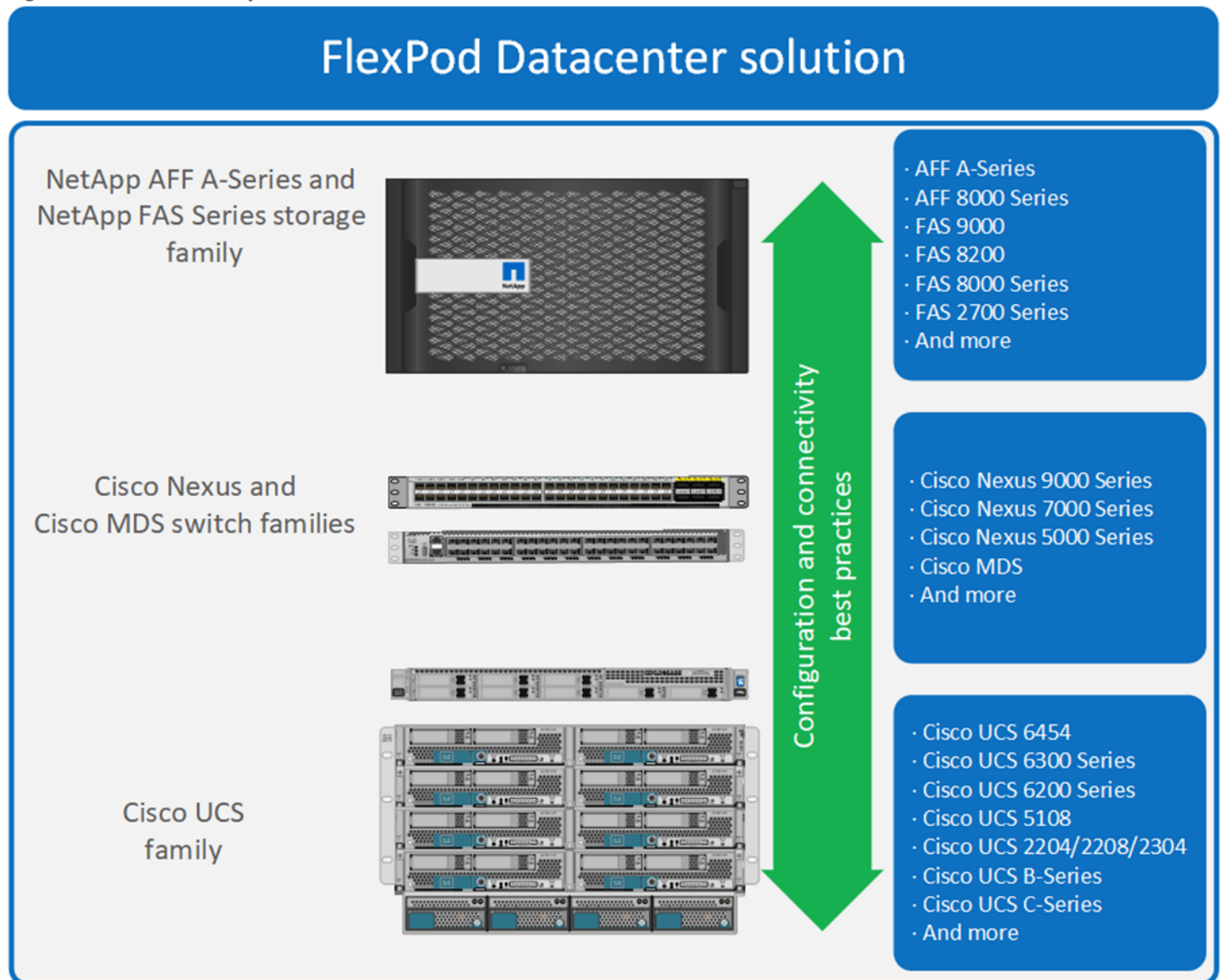
FlexPod Overview

FlexPod is a best-practices data center architecture that includes the following components (Figure 1):

- Cisco Unified Computing System™ (Cisco UCS®)
- Cisco Nexus® switches
- Cisco® MDS switches
- NetApp All Flash FAS (AFF) systems

Figure 1.   FlexPod component families

These components are connected and configured according to the best practices of both Cisco and NetApp to provide an excellent platform for running a variety of enterprise workloads with confidence. FlexPod can scale up for greater performance and capacity (adding computing, network, and storage resources individually as needed), or it can scale out for environments that require multiple consistent deployments (for instance, rollout of additional FlexPod stacks). The reference architecture discussed in this document uses Cisco Nexus 9000 Series Switches for the network switching element.

One of the main benefits of FlexPod is its ability to maintain consistency with scale. Each of the component families shown in Figure 1 (Cisco UCS, Cisco Nexus, Cisco MDS, and NetApp AFF) offers platform and resource options to scale the infrastructure up or down, while supporting the same features and functions that are required under the configuration and connectivity best practices of FlexPod.

FlexPod design principles

FlexPod addresses four main design principles: availability, scalability, flexibility, and manageability. The architecture goals are as follows:

- **Application availability:** Helps ensure that services are accessible and ready to use

- **Scalability:** Addresses increasing demands with appropriate resources

- **Flexibility:** Provides new services and recovers resources without requiring infrastructure modification

- **Manageability:** Facilitates efficient infrastructure operations through open standards and APIs

FlexPod design for NetApp Memory-Accelerated Data

The Cisco Nexus 9000 Series Switches support two modes of operation: NX-OS standalone mode, using Cisco NX-OS Software, and ACI fabric mode, using the Cisco Application Centric Infrastructure (Cisco ACI™) platform. In standalone mode, the switch performs like a typical Cisco Nexus switch, with increased port density, low latency, and 100, 40, 10, and 25 Gigabit Ethernet connectivity. In fabric mode, the administrator can take advantage of the Cisco ACI platform. The design discussed here uses the standalone mode.

FlexPod with NX-OS is designed to be fully redundant in the computing, network, and storage layers. There is no single point of failure from a device or traffic-path perspective. Figure 2 shows the connection of the various elements of the FlexPod design for NetApp Memory-Accelerated Data (MAX Data).

Figure 2.   NetApp MAX Data for FlexPod lab topology
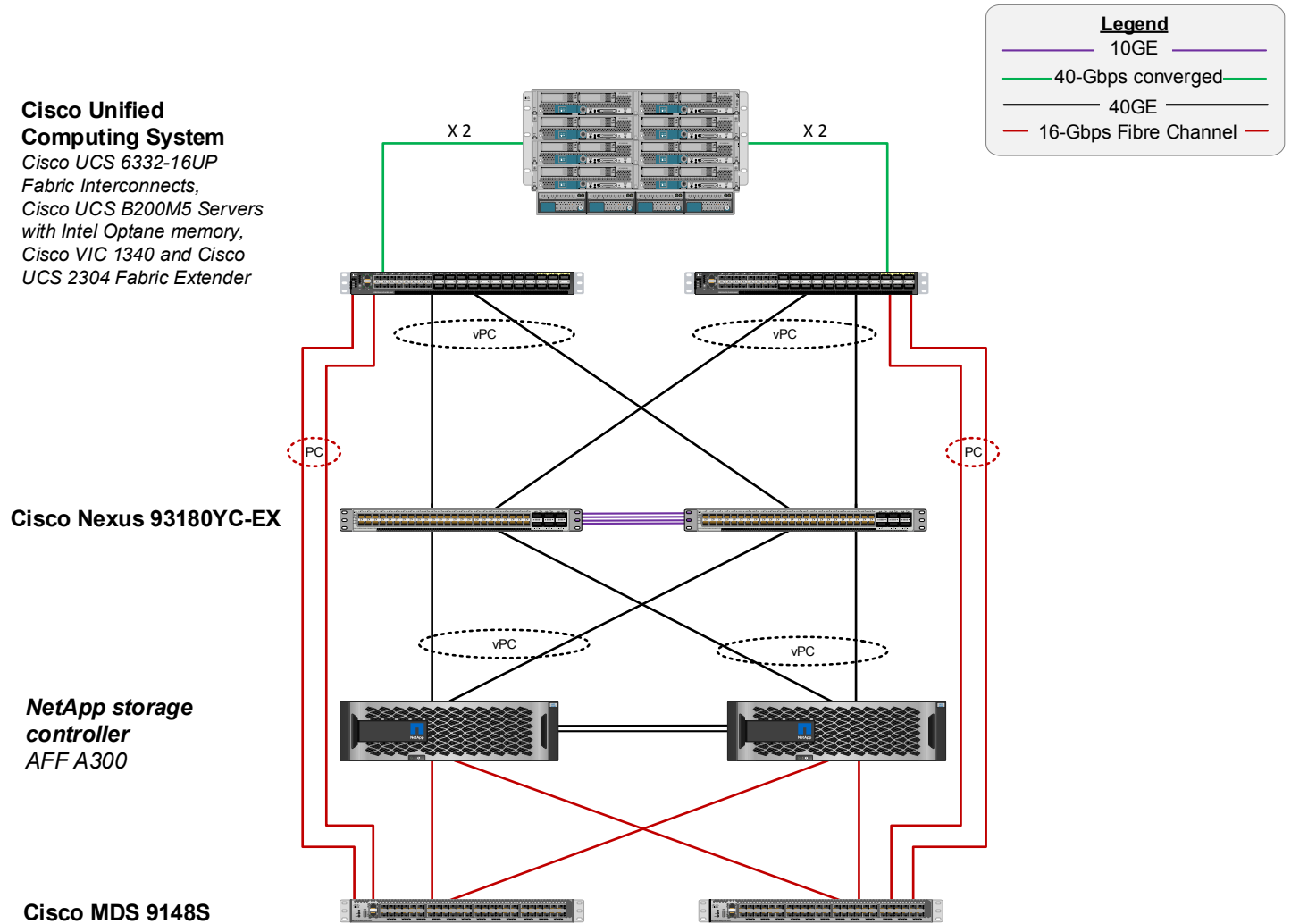
This design uses the third-generation Cisco UCS 6332-16UP Fabric Interconnects and the Cisco UCS Virtual Interface Card (VIC) 1300 platform in the servers. The Cisco UCS B200 M5 Blade Servers in the Cisco UCS chassis use the Cisco UCS VIC 1340 with Port Expander connected to the Cisco UCS 2304 Fabric Extender I/O module (IOM), and each virtual network interface card (vNIC) has a speed of 40 Gbps. The fabric interconnects connect through 40-Gbps port channels to virtual port channels (vPCs) across the Cisco Nexus 93180YC-EX Switches.

The connectivity between the Cisco Nexus switches and the NetApp AFF A300 storage cluster is also 40 Gbps with port channels on the storage controllers and vPCs on the switches.

This configuration supports IP-based storage protocols (Network File System [NFS], Common Internet File System [CIFS], and Small Computer System over IP [iSCSI]) over a high-speed network between the storage and the Cisco UCS servers.

For Fibre Channel SAN connectivity, this design also uses the third-generation Cisco UCS 6332-16UP Fabric Interconnects and the Cisco UCS VIC 1300 platform in the servers. The Cisco UCS B200 M5 Blade Servers in the Cisco UCS chassis use the Cisco UCS VIC 1340 with Port Expander connected to the Cisco UCS 2304 Fabric Extender IOM, and each Fibre Channel over Ethernet (FCoE) virtual host bus adapter (vHBA) has a speed of 40 Gbps.

The fabric interconnects connect through 16-Gbps SAN port channels to the Cisco MDS 9148S Fibre Channel switches.

The connectivity between the Cisco MDS switches and the NetApp AFF A300 storage cluster is also 16-Gbps Fibre Channel.

This configuration supports 16-Gbps Fibre Channel, for Fibre Channel Protocol (FCP) between the storage cluster and Cisco UCS.

Intel Optane DC persistent memory in FlexPod

Intel Optane DC persistent memory is a new memory technology based on 3D XPoint technology. It is nonvolatile, higher density than DRAM, and faster than NAND technology with super low latency and very high endurance. 3D XPoint architecture is different from DRAM in that it interleaves microscopic charging wires between stacked memory and selector cells. The Intel Optane Data Center Persistent Memory Module (DCPMM) can be used in three operating modes and must be paired with DRAM DIMMs in each memory channel (see the Cisco memory guide listed in the reference section at the end of this document):

- It can be used in memory mode, which provides additional volatile memory capacity.
- It can be used in App Direct mode, which provides persistence and can be accessed through traditional memory load and store operations.
- It can be used in mixed mode where a percentage of DCPMM is used in memory mode and the rest in App Direct mode.

Intel initially shipped Optane DCPMM with a capacity of 128 GB, with capacities of 256 and 512 GB per module to follow. Optane DCPMM 128-GB DIMMs were released on April 2, 2019. In this lab validation, Optane DCPMM 256-GB DIMMs were used.

NetApp MAX Data in FlexPod

NetApp MAX Data is an application acceleration product that uses the latest server-side storage technologies to improve application response time and provide enterprise-class data protection to persistent memory. MAX Data provides a file system for your application data with the performance of Intel Optane DCPMM and the capacity of a NetApp AFF storage array (Figure 3). MAX FS, the file system for MAX Data, spans the memory and storage tiers in a single file system and intelligently moves (tiers) data between the memory tier and the storage tier based on use. It is not a cache; its storage in persistent local memory keeps the hottest, or most frequently accessed, data in the memory tier, and tiers the cooler, or less frequently used, data to the storage tier. This behavior allows you to accelerate your application working set at memory speed while allowing a much larger total data-set size. MAX Data also provides enterprise data services to Optane DCPMM with snapshots and replication technologies. NetApp, as an industry leader, is the first to offer an application-independent solution that merges the highest local speeds with global data protection.

Figure 3. NetApp MAX Data stack
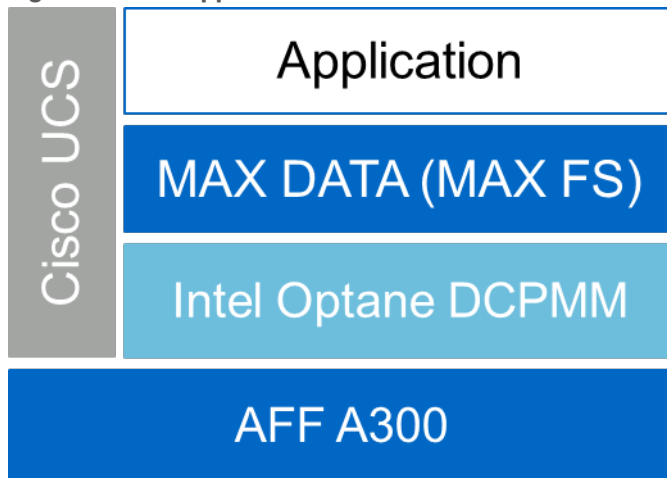
Application use cases for NetApp MAX Data and Intel Optane DCPMM

MAX Data can be used with almost any database or application. It can be used with relational databases such as MySQL, and PostgreSQL. It can be used with NoSQL databases such as MongoDB, Cassandra, Couchbase, and Redis. It can even be used with custom applications that require lower latency. These databases are used in many different scenarios and often benefit greatly from reduced application response time. Some of the most popular use cases today in which reduced application response time is crucial are financial trading, fraud and cyber-threat detection, healthcare diagnostics and genomic research, E-commerce, artificial intelligence (AI), and the Internet of Things (IoT). In each case, microseconds matter. Being able to make decisions faster is critical to gaining competitive advantage–or saving a life.

Implementation

Intel Optane DCPMM can be used in many scenarios as noted in the preceding section. MAX Data uses Optane DCPMM as persistent memory in App Direct mode. This is the way (in VMware ESXi and Red Hat Enterprise Linux [RHEL] 7.6) to use Optane DCPMM in a persistent manner and is the reason that application vendors must rewrite their applications to use Intel Optane DCPMM. With MAX Data, customers can use Optane DCPMM without the need to rewrite their applications, nor do the database or application vendors need to support Device Direct Access mode. It makes Optane DCPMM plug-and-play easy.

Here are the high-level steps to using Optane DCPMM with MAX Data in your application servers today:

1. Install Optane DCPMM in the Cisco UCS B200 M5 Blade Server.

2. Configure the persistent memory policy of the service profile policy and modify the goal to use 0 percent memory mode.

3. Install vSphere 6.7U2, RHEL, or CentOS 7.6.

4. For RHEL or CentOS, create two namespaces (one per CPU); for VMware, add a NVDIMM to the guest virtual machine.

5. Install and configure MAX Data.

6. Use the file system with your application as you would a traditional block storage device.

Performance

MAX Data performance reduces the latency for critical I/O operations from a millisecond or more to around 10 microseconds: more than an order of magnitude improvement. Even over the fastest shared storage systems with the lowest latency, MAX Data can provide substantial performance improvements because the working set of the data set resides in the memory tier, right next to the CPU. With this reduced latency, databases can support more transactions using fewer computing resources and complete user queries faster.

Data protection

MAX Data brings enterprise-class data protection to Optane DCPMM. No database solution is complete without data protection. MAX Data integrates seamlessly with NetApp ONTAP, storing consolidated snapshot-based backups of your database, enabling quick and reliable backup and recovery, including comprehensive disaster recovery with NetApp SnapMirror. MAX Data software is flexible and can be deployed in various configurations to match application needs. For the most demanding recovery requirements, you can use NetApp MAX Recovery. MAX Recovery replicates every MAX Data transaction on the application server so that if the application server fails the data can be recovered. MAX Recovery in conjunction with ONTAP allows rapid recovery, up to the last transaction, if a database server suddenly fails.

Supported configurations

Consult the NetApp Interoperability Matrix Tool for up-to-date information about supported configurations.

- Memory technologies supported: DRAM, NVDIMM-N, and Intel Optane DCPMM
- Operating systems supported: VMware vSphere 6.7U2 or later (for Optane DCPMM Support) and RHEL and CentOS 7.5 or later (RHEL and CentOS 7.6 for Optane DCPMM support)
- Storage platforms supported: NetApp AFF A-Series and AFF8000 Series
- Storage protocols supported: Fibre Channel and iSCSI
- Storage operating systems supported: NetApp ONTAP 9.5 or later

System validation and testing (bare metal)

This section provides a high-level summary of the MAX Data on FlexPod validation testing. The solution was validated for basic data access and performance by deploying the flexible I/O (fio) tester tool on the RHEL host.

This FlexPod configuration was successfully verified by performing multiple tests to determine persistence and performance:

- Write data to MAX FS and reboot the system to validate Optane DCPMM persistence of the data in the memory tier.
- Use fio to read and write data and increase the number of threads to concurrency.

Validated hardware and software

Table 1 lists the hardware and software versions used during the solution validation process. Note that Cisco and NetApp have interoperability matrixes that should be referenced to determine support for any specific implementation of FlexPod. See the following documents for more information:

- NetApp Interoperability Matrix Tool
- Cisco UCS Hardware and Software Interoperability Tool

Table 1.   Validated hardware and software versions (bare metal)

Layer	Device	Image	Comments
Computing	Cisco UCS 6332-16UP Fabric Interconnect and Cisco UCS B200 M5 Blade Server with Cisco UCS VIC 1340	Release 4.0(4b)	
CPU	Intel Xeon Platinum 8280 Second-Generation Scalable processor		
Memory	12 x 32-GB DRAM 4 x 256-GB Intel Optane DCPMM		
Network	Cisco Nexus 93180YC-EX Switch in NX-OS standalone mode	Release 7.0(3)I7(6)	
Storage network	Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch	Release 8.3(2)	
Storage	NetApp AFF A300	NetApp ONTAP 9.5P4	
Software	RHEL 7.6	Kernel 3.10.0-957	
Software	NetApp MAX Data 1.3		
	Cisco UCS VIC fnic driver	Release 2.0.0.42-77.0	
	Cisco UCS VIC enic driver	Release 3.2.210.18-738.12	

Bare metal test results

The test setup for this solution used one Cisco UCS B200 M5 server with two Intel Second-Generation Scalable processors and four 256-GB Intel Optane DCPMM DIMMs with Cisco UCS VIC 1340 adapters running the RHEL 7.6 OS with NetApp MAX Data 1.3 and fio. Multiple tests were run with fio iterating on the number of threads (1, 2, 4, 8, 16, and 20), using a 4-KB block size with 100 percent random write and random read operations. I/O operations per second (IOPS) and latency were measured with and without MAX Data. An example of the fio string used in the validation of this configuration is shown here:

```
fio --bs=4K --rw=randwrite --direct=1 -numjobs=20 --size=10g --ioengine=psync --
norandommap --time_based --runtime=120 --directory=/mnt/mount --group_reporting --
name=fio_test --output-format=normal -output=outfile.txt
```

The graphs in Figures 4 and 5 demonstrate that for the fio workload used in this validation testing, MAX Data on FlexPod is capable of 5 times more I/O operations with 25 times less latency. If your applications

need to do more work and to do this work faster, you can use MAX Data on FlexPod to help deliver the performance you require, supporting the most critical enterprise applications.

Figure 4.   Performance benefits of adding NetApp MAX Data to NetApp A300
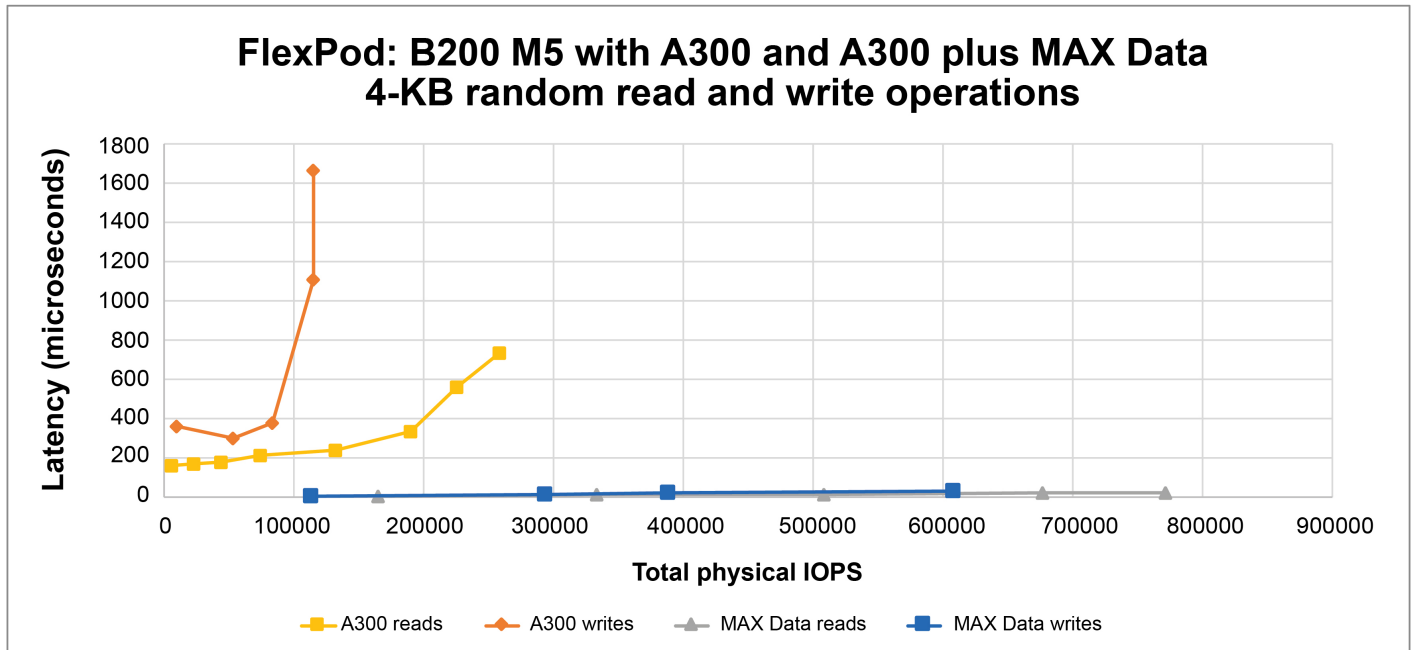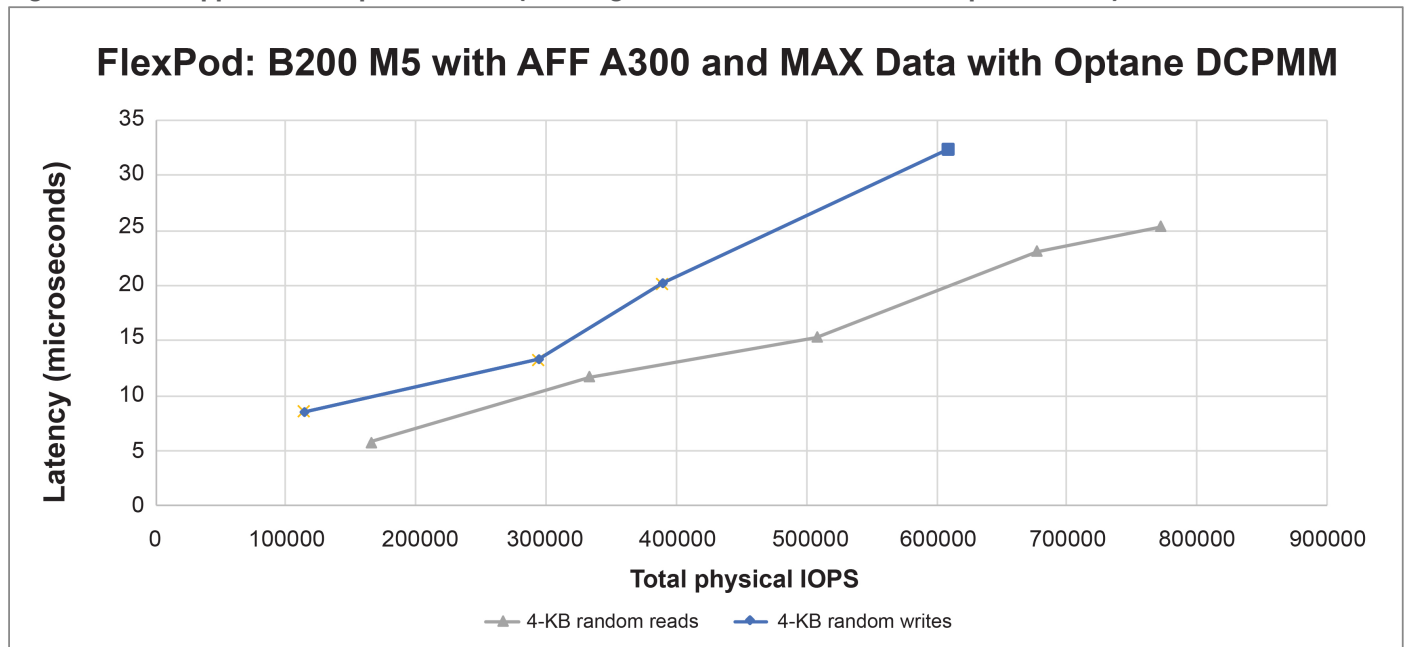

Figure 5.   NetApp MAX Data performance (showing MAX Data's microsecond response times)

System validation and testing (VMware)

This section provides a high-level summary of the MAX Data on FlexPod validation testing. The solution was validated for basic data access and performance by deploying the flexible I/O (fio) tester tool on the RHEL guest.

This FlexPod configuration was successfully verified by performing multiple tests to determine persistence and performance:

- Write data to MAX FS and reboot the system to validate Optane DCPMM persistence of the data in the memory tier.

- Use fio to read and write data and increase the number of threads to concurrency.

Validated hardware and software

Table 2 lists the hardware and software versions used during the solution validation process. Note that Cisco and NetApp have interoperability matrixes that should be referenced to determine support for any specific implementation of FlexPod. See the following documents for more information:

- NetApp Interoperability Matrix Tool

- Cisco UCS Hardware and Software Interoperability Tool

Table 2. Validated hardware and software versions (VMware Hypervisor)

| Layer	Device	Image	Comments
Computing	Cisco UCS 6332-16UP Fabric Interconnect and Cisco UCS B200 M5 Blade Server with Cisco UCS VIC 1340	Release 4.0(4b)	
CPU	Intel Xeon Platinum 8280 Second-Generation Scalable processor		
Memory	12 x 32-GB DRAM 4 x 256-GB Intel Optane DCPMM		
Network	Cisco Nexus 93180YC-EX Switch in NX-OS standalone mode	Release 7.0(3)I7(6)	
Storage network	Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch	Release 8.3(2)	
Storage	NetApp AFF A300	NetApp ONTAP 9.5P4	
Hypervisor	VMware vSphere / ESXi	6.7 Update 2	
Guest operating system	RHEL 7.6	Kernel 3.10.0-957	
Software	NetApp MAX Data 1.3		
	Cisco UCS VIC fnic driver	Release 4.0.0.35	
	Cisco UCS VIC enic driver	Release 1.0.29.0	

VMware guest test results

The test setup for this solution used one Cisco UCS B200 M5 server with two Intel Second-Generation Scalable processors and four 256-GB Intel Optane DCPMM DIMMs with Cisco UCS VIC 1340 adapters running VMware ESXi 6.7 Update 2 with a guest virtual machine running the RHEL 7.6 OS with NetApp MAX Data 1.3 and fio. An NVDIMM of 990 GB was added to the virtual machine and a namespace automatically created by VMware. Multiple tests were run with fio iterating on the number of threads (1, 2, 4, 8, 16, and 20), using a 4-KB block size with 100 percent random writes operations. IOPS and latency were measured with and without MAX Data. An example of the fio string used in the validation of this configuration is shown here:

```
fio --bs=4K --rw=randwrite --direct=1 –numjobs=20 --size=10g --ioengine=psync --
norandommap --time_based --runtime=120 --directory=/mnt/mount --group_reporting --
name=fio_test --output-format=normal –output=outfile.txt
```

Figure 6 demonstrates that for the fio workload used in this validation testing MAX Data on FlexPod is capable of 2.5 times more I/O operations with 12.5 times less latency. If your applications need to do more work and to do this work faster, you can use MAX Data on FlexPod to help deliver the performance you require, supporting the most critical enterprise applications.

Figure 6.   Performance benefits of adding NetApp MAX Data to NetApp A300 running on a RHEL VM on VMware
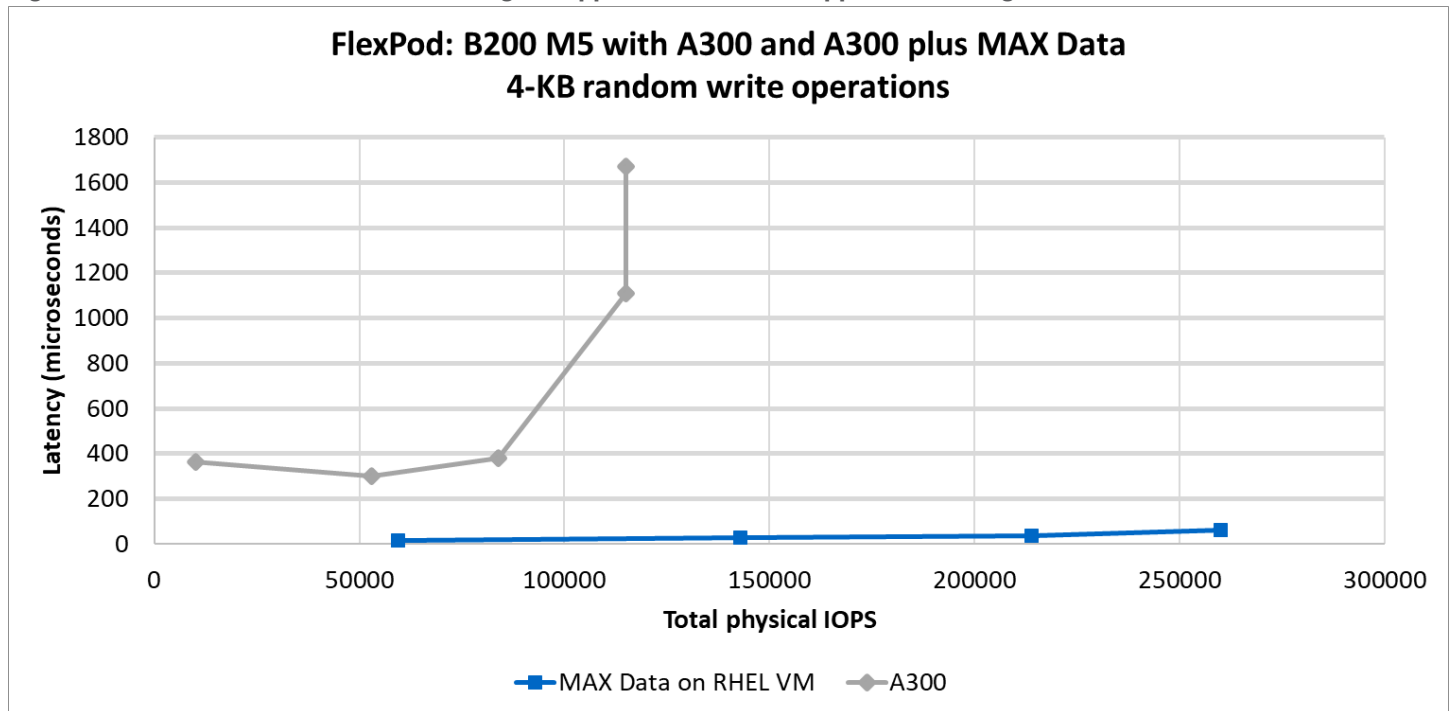
PostgreSQL performance validation

This solution validation demonstrates that MAX Data can accelerates databases such as PostgreSQL. The solution employs NetApp MAX Data software running on FlexPod with Optane DCPMM. Moreover, MAX Data is so efficient in its CPU and DRAM usage that users can serve more transactions per minute, at a better response time.

In this validation PostgreSQL was installed on a Red Hat Linux virtual machine running on VMware vSphere 6.7 Update 2. These tests were performed using XFS and then using MAX Data. For the XFS virtual

machine, 24 vCPUs and 80 GB of DRAM were used. For the MAX Data virtual machine, 12 vCPUs and 40 GB of DRAM and 180 GB of Optane DCPMM were provisioned to the virtual machine. On the MAX Data virtual machine, MAX Data 1.3 was installed and configured, and a PostgreSQL database was created and populated. Both tests used 4 LUNs for the DB and 4 LUNS for the Log files.

In this test DBT2 benchmark was used to generate a TPC-C like load. DBT2 is an OLTP transactional performance test. It simulates a wholesale parts supplier where several workers access a database, update customer information, and check on parts inventories (Figure 7). The diagram below illustrates the hardware and virtual machine configuration differences between the XFS and MAX Data tests.

Figure 7.   PostgreSQL validation hardware and software stack
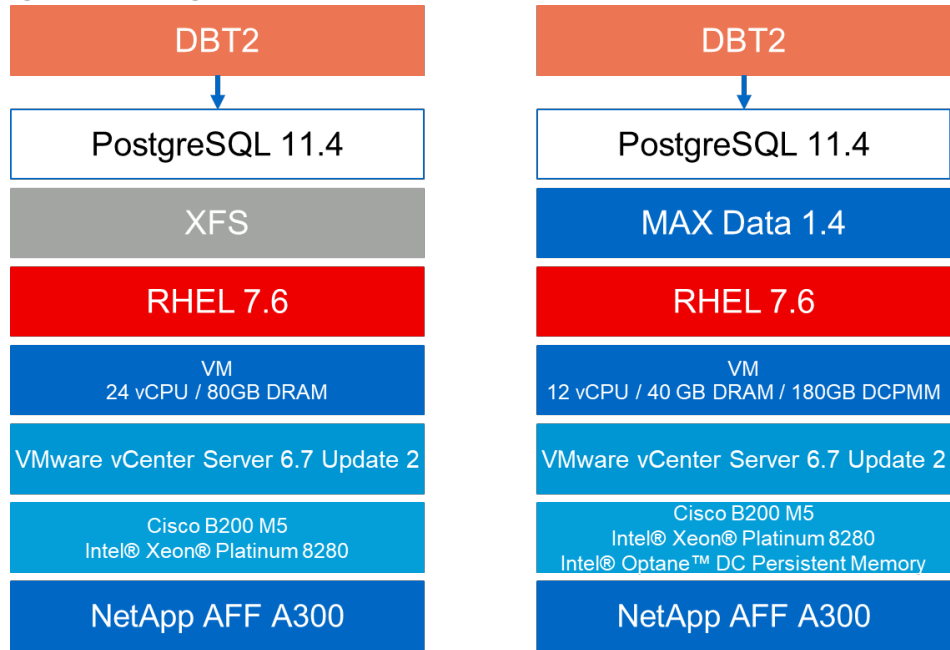
Validated hardware and software

Table 3 lists the hardware and software versions used during the PostgreSQL solution validation process. Note that Cisco and NetApp have interoperability matrixes that should be referenced to determine support for any specific implementation of FlexPod. See the following documents for more information:

- NetApp Interoperability Matrix Tool
- Cisco UCS Hardware and Software Interoperability Tool

Table 3.   Validated hardware and software versions (PostgreSQL test)

Layer	Device	Image	Comments
Computing	Cisco UCS 6332-16UP Fabric Interconnect and Cisco UCS B200 M5 Blade Server with Cisco UCS VIC 1340	Release 4.0(4b)	
CPU	Intel Xeon Platinum 8280 Second-Generation Scalable processor		
Memory	12 x 32-GB DRAM 4 x 256-GB Intel Optane DCPMM		
Network	Cisco Nexus 93180YC-EX Switch in NX-OS standalone mode	Release 7.0(3)I7(6)	
Storage network	Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch	Release 8.3(2)	
Storage	NetApp AFF A300	NetApp ONTAP 9.5P4	
Hypervisor	VMware vSphere / ESXi	6.7 Update 2	
Guest operating system	RHEL 7.6	Kernel 3.10.0-957	
Software	PostgreSQL	11.4	
	NetApp MAX Data 1.3		
	Cisco UCS VIC fnic driver	Release 4.0.0.35	
	Cisco UCS VIC enic driver	Release 1.0.29.0	

Software configuration

To keep the comparison fair, both virtual machines use the same hypervisor and operating system. PostgreSQL, RHEL, and hardware versions are all new as of 2019. However, the CPU and memory were reduced by half from the test with XFS to the MAX Data.

Following are example command lines for running the benchmarks, regardless of whether that data landed on XFS or MAX Data:

`postgresql.conf file`

```
Shared buffers=20gb

max_connections=300

max_wal_size=15gb

max_files_per_process=10000

wal_sync_method= open_sync

checkpoint_completion_target=0.9
```

DBT2 client

```
Load phase: dbt2-pgsql-build-db -r -w {numofwarehouses} (scale=3200 for a 320GB DB
+ 15GB WAL)

Run phase: dbt2-run-workload -a pgsql -D dbt2 -H {host_ip} -l 5432 -d 3600 -w
{warehouses} -o /home/postgres/results -c 32 -t 1 -s 5
```

PostgreSQL performance test Results

The validation demonstrates a significant improvement in performance. During the validation the number of transactions increased by 2.4 times per minute and at the same time reducing database response time by 2 to 2.4 times all while reducing the virtual hardware. The next sections and graphs demonstrate this increase in performance and the reduction of hardware.

PostgreSQL increase in transactions per minute

DBT2 benchmark was used to generate a TPC-C like load. DBT2 is an OLTP transactional performance test. It simulates a wholesale parts supplier where several workers access a database, update customer information, and check on parts inventories (Figure 8).

Figure 8. Transactions per minute for DBT2 when the database size is twice the size of Optane. Note that MAX Data throughput and gap compared to XFS is even greater for smaller databases.
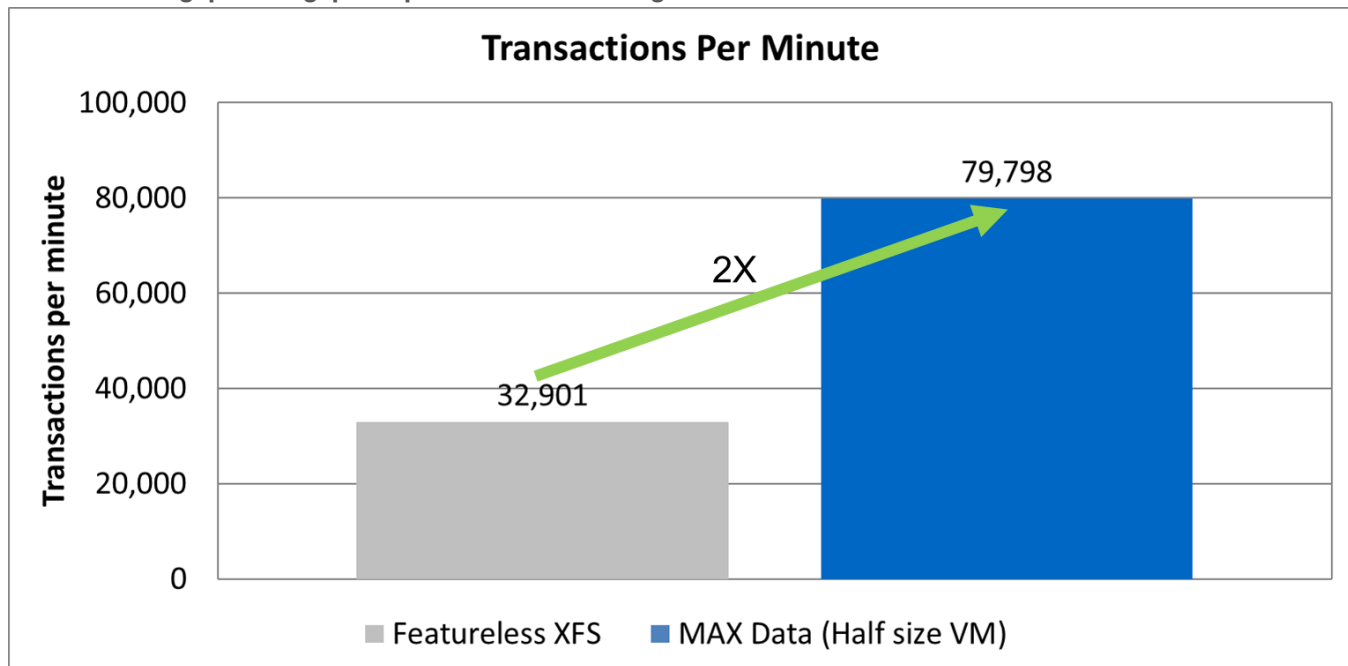

Figure 8 reveals that even with half of the CPUs and DRAM budget, MAX Data using Optane PM achieved 2.43 times the throughput improvement over the more expensive baseline virtual machine.

PostgreSQL reducing end-user response time

Virtual machine density and database throughput typically appeal to the administrator and CIO. Users, however, mostly care about the response time. Figure 9 reveals that on top of the better throughput with fewer resources, MAX Data also improves latency. Response time was measured to improve by 2.4 and 2 times for average and 90 percentile latencies, respectively.

Figure 9. PostgreSQL response time with and without MAX Data
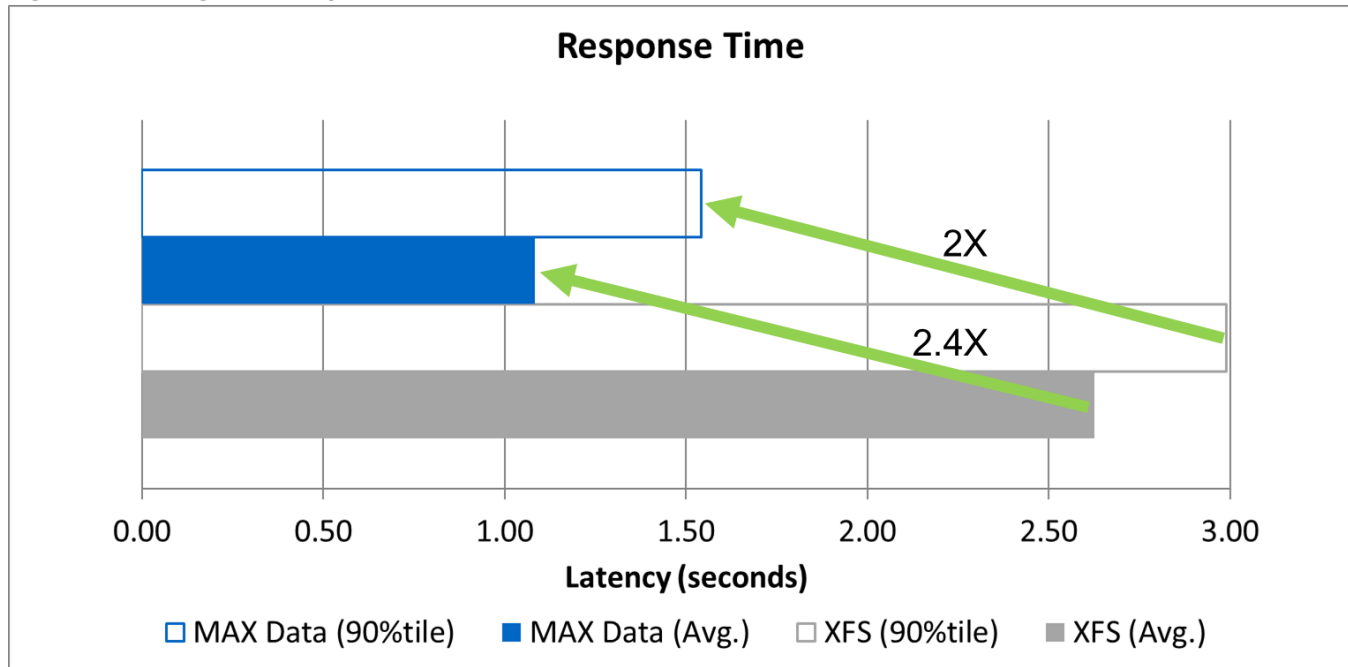
MAX data efficiency

As stated earlier, MAX Data is efficient in its CPU usage. Table 4 shows the average CPU usage during the XFS and MAX Data testing. Testing results show that the number of both CPUs and vCPUs from the virtual machine were reduced. Because the application was not waiting for I/O, more transactions were completed with fewer resources.

Table 4. CPU efficiency of MAX Data

File system	vCPU cores	%user	%system	%iowait	%idle
XFS	24 vCPUs	4.66	6.86	47.8	40.67
MAX Data	12 vCPU	31.37	17.4	0.01	51.22

Conclusion

FlexPod Datacenter is the optimal shared infrastructure foundation for NetApp MAX Data to allow high-performance access to applications that need it. As Intel Optane DCPMM and MAX Data capacity, performance, and data-protection capabilities increase, FlexPod is well suited as the platform of choice, providing the scalability and reliability needed to support these capabilities.

With FlexPod, Cisco and NetApp have created a platform that is both flexible and scalable for multiple use cases and applications. With MAX Data and Optane DCPMM, FlexPod adds yet another feature to help organizations efficiently and effectively support business-critical applications running simultaneously on the same shared infrastructure. The flexibility and scalability of FlexPod enables customers to start with a right-sized infrastructure that can grow with and adapt to their evolving business requirements.

For more information

Consult the following references for additional information about the topics discussed in this document.

Products and solutions

- Cisco Unified Computing System:
  http://www.cisco.com/en/US/products/ps10265/index.html

- Cisco UCS 6300 Series Fabric Interconnects:
  https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-6300-series-fabric-interconnects/index.html

- Cisco UCS 5100 Series Blade Server Chassis:
  http://www.cisco.com/en/US/products/ps10279/index.html

- Cisco UCS B-Series Blade Servers:
  http://www.cisco.com/en/US/partner/products/ps10280/index.html

- Cisco UCS C-Series Rack Servers:
  http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html

- Cisco UCS adapters:
  http://www.cisco.com/en/US/products/ps10277/prod_module_series_home.html

- Cisco UCS Manager:
  http://www.cisco.com/en/US/products/ps10281/index.html

- Intel Optane DC Persistent Memory Module:

- https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html

- Cisco Nexus 9000 Series Switches:
  http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html

- NetApp ONTAP 9:
  http://www.netapp.com/us/products/platform-os/ontap/index.aspx

- NetApp AFF A-Series:

- http://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx

- NetApp MAX Data:
  https://www.netapp.com/us/products/data-management-software/max-data.aspx

Interoperability matrixes

- Cisco UCS Hardware Compatibility Matrix:
  https://ucshcltool.cloudapps.cisco.com/public/

- NetApp Interoperability Matrix Tool:
  http://support.netapp.com/matrix/

Configuration guides

- Cisco memory guide: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/memory-guide-c220-c240-b200-m5.pdf

Printed in USA

219245-8   08/19