

White Paper Cisco Public

Introducing End-to-End NVMe for FlexPod

Contents

FlexPod overview	3
NVMe storage and end-to-end FC-NVMe in FlexPod	6
System validation and testing	11
Conclusion	16
For more information	16

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



FlexPod Datacenter solution

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, or storage resources individually as needed), or it can scale out for environments that require multiple consistent deployments (such as rolling out 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, 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

Latest FlexPod design

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 40 and 100 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 latest FlexPod design used in this validation of FC-NVMe.



Figure 2.

Latest FlexPod design used in this validation

From a Fibre Channel SAN perspective, this design uses the latest fourth-generation Cisco UCS 6454 Fabric Interconnects and the Cisco UCS VICs 1400 platform in the servers. The Cisco UCS B200 M5 Blade Servers in the Cisco UCS chassis use the Cisco UCS VIC 1440 connected to the Cisco UCS 2208 Fabric Extender IOM, and each Fibre Channel over Ethernet (FCoE) virtual host bus adapter (vHBA) has a speed of 20 Gbps. The Cisco UCS C220 M5 Rack Servers managed by Cisco UCS use the Cisco UCS VIC 1457 with two 25-Gbps interfaces to each fabric interconnect. Each C220 M5 FCoE vHBA has a speed of 50 Gbps.

The fabric interconnects connect through 32-Gbps SAN port channels to the latest-generation Cisco MDS 9148T or 9132T Fibre Channel switches.

The connectivity between the Cisco MDS switches and the NetApp AFF A800 storage cluster is also 32-Gbps Fibre Channel.

This configuration supports 32-Gbps Fibre Channel, for Fibre Channel Protocol (FCP), and NVMe over Fibre Channel (FC-NVMe, explained in the next section of this document) storage between the storage cluster and Cisco UCS. For this validation, 4 Fibre Channel connections to each storage controller are used. On each storage controller, the 4 Fiber channel ports are used for both FCP and FC-NVME protocols.

From an IP perspective, this design also uses the latest fourth-generation Cisco UCS 6454 Fabric Interconnects and the Cisco UCS Virtual Interface Card (VIC) 1400 platform in the servers. The Cisco UCS B200 M5 Blade Servers in the Cisco UCS chassis use the Cisco UCS VIC 1440 connected to the Cisco UCS 2208 Fabric Extender I/O module (IOM), and each virtual network interface card (vNIC) has a speed of 20 Gbps. The Cisco UCS C220 M5 Rack Servers managed by Cisco UCS use the Cisco UCS VIC 1457 with two 25-Gbps interfaces to each fabric interconnect. Each C220 M5 vNIC has a speed of 50 Gbps. The fabric interconnects connect through 100-Gbps port channels to virtual port channels (vPCs) across the latest-generation Cisco Nexus 9336C-FX2 Switches.

The connectivity between the Cisco Nexus switches and the latest-generation NetApp AFF A800 storage cluster is also 100 Gbps with port channels on the storage controllers and vPCs on the switches. The NetApp AFF A800 storage controllers are equipped with Non-Volatile Memory Express (NVMe) disks on the higher-speed Peripheral Connect Interface Express (PCIe) bus.

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.

NVMe storage and end-to-end FC-NVMe in FlexPod

The NVMe data storage standard is emerging as a core technology. NVMe is transforming enterprise data storage access and transport by delivering very high bandwidth and very low latency storage access for current and future memory technologies. NVMe replaces the SCSI command set with the NVMe command set. It relies on PCIe, a high-speed and high-bandwidth hardware protocol that is substantially faster than older standards like SCSI, SAS, and SATA.

NVMe was designed to work with nonvolatile flash drives, multicore CPUs, and gigabytes of memory. It also takes advantage of the significant advances in computer science since the 1970s, enabling streamlined command sets that more efficiently parse and manipulate data.

An end-to-end NVMe architecture also enables data center administrators to rethink the extent to which they can push their virtualized and containerized environments and the amount of scalability that their transaction-oriented databases can support. It seamlessly extends a customer's existing SAN infrastructure for real-time applications while simultaneously delivering improved IOPS and throughput with reduced latency.

NVMe reduces the cost of IT by efficiently using all resources throughout the stack. An NVMe over Fibre Channel solution is lossless and can handle the scalability requirements of next-generation applications. These new technologies include artificial intelligence, machine learning, deep learning, real-time analytics, and other mission-critical applications.

FlexPod is the ideal platform for introducing FC-NVMe. It can be supported with the addition of the Cisco UCS VIC 1400 series in existing Cisco UCS B200 M5 servers and simple, nondisruptive software upgrades to the Cisco UCS system, the Cisco MDS 32-Gbps switches, and the NetApp AFF storage arrays. Once the supported hardware and software are in place, the configuration of FC-NVMe is similar to FCP configuration (details to follow).

NetApp and FC-NVMe

NetApp ONTAP 9.5 onwards provides a complete NVMe over Fibre Channel, or FC-NVMe, solution. A nondisruptive ONTAP software update for AFF A300, AFF A400, AFF A700, AFF A700s, and AFF A800 arrays allows these devices to support an end-to-end NVMe storage stack. Therefore, servers with sixth-generation host bus adapters (HBAs) and NVMe driver support can communicate with these arrays using native NVMe.

NVMe is causing an architectural shift that makes communication with storage systems massively parallel. The result is greater bandwidth and lower-latency connectivity between servers and storage devices.

Many other factors enable the NetApp implementation of NVMe to provide exceptional performance in the data center, including the following:

- Interrupt handling
- Internal locking needed to serialize I/O requests
- Command-set streamlining
- Reduced context switches
- Lockless design
- Polling mode

The following features and limitations characterize the ONTAP implementation of FC-NVMe:

- NVMe/FC relies on the ANA protocol to provide multipathing and path management necessary for both path and target failover. The ANA protocol defines how the NVMe subsystem communicates path and subsystem errors back to the host so that the host can manage paths and failover from one path to another. ANA fills the same role in NVMe/FC that ALUA does for both FCP and iSCSI protocols. ANA with host OS path management such as NVME-Multipath or DM-Multipath provide path management and failover capabilities for NVMe/FC. SUSE Linux Enterprise Server (SLES) 15 features Asymmetric namespace access (ANA) support with ONTAP 9.6.
- FC-NVMe is supported on 32-Gbps end-to-end configurations from ONTAP 9.6 (or above).
- ONTAP 9.6 supports maximum 4 FC-NVMe logical interfaces (LIFs) per Storage VM.
- The SLES 15 initiator host can serve both FC-NVMe and Fibre Channel over SCSI (FC-SCSI) traffic through the same 32-Gbps Fibre Channel adapter ports. Indeed, this is expected to be the most commonly deployed host configuration for end users. For FCP (FC-SCSI), you can configure dmmultipath as usual for the SCSI logical unit numbers (LUNs), or logical devices, resulting in multipath devices, whereas NVME multipath is used to configure FC-NVME multipath namespace devices (i.e /dev/nvmeXnY) on the SLES15 initiator host.

Cisco UCS and FC-NVMe

Cisco UCS provides a unified fabric that is an architectural approach delivering flexibility, scalability, intelligence and simplicity. This flexibility allows Cisco UCS to readily support new technologies such as FC-NVMe seamlessly. In Cisco UCS software release 4.0(2), support for FC-NVMe, which defines a mapping protocol for applying the NVMe interface to Fibre Channel was introduced. This release adds support for the FC NVME Initiator adapter policy on UCS 6300 Series Fabric Interconnects and UCS 6454 Fabric Interconnects. Cisco UCS software release 4.0(4) is used for validating and finding the performance characteristics in this paper.

Cisco UCS Software Release 4.0(4), along with the Cisco UCS VIC 1400 platform adapters and the Release 2.0.0.42-77.0 Fibre Channel network interface card (fnic) driver, supports FC-NVMe. In a Cisco UCS service profile, both standard Fibre Channel and FC-NVMe vHBAs can be created. The type of vHBA is selected in the Fibre Channel adapter policy, as shown in Figure 3.

vHBA Type

○ FC Initiator ○ FC Target ● FC NVME Initiator ○ FC NVME Target

Figure 3.

FC-NVMe vHBA type selection

A default Fibre Channel adapter policy named FCNVMeInitiator is preconfigured in Cisco UCS Manager. This policy contains recommended adapter settings for FC-NVMe. Since SLES 15 was the server operating system used in this validation, the default Fibre Channel adapter policy named Linux was used for Fibre Channel vHBAs.

Both Fibre Channel and FC-NVMe vHBAs can exist in a Cisco UCS service profile on a single server. In the lab validation for this document, four vHBAs (one FC-NVME initiator on each Fibre Channel fabric and one Fibre Channel initiator on each Fibre Channel fabric) were created in each service profile to facilitate performance comparisons between the two protocols. Each vHBA, regardless of type, was automatically assigned a worldwide node name (WWNN) and a worldwide port name (WWPN). The Cisco UCS fabric interconnects were in Fibre Channel end-host mode (NPV mode) and uplinked through a SAN port channel to the Cisco MDS 9132T switches in NPV mode. Zoning in the Cisco MDS 9132T switches connected the vHBAs to storage targets for both FC-NVMe and Fibre Channel. Single-initiator, multiple-target zones were used for both FCP and FC-NVMe. The connectivity to the NVMe subsystems on storage is configured in both the storage and the host operating system.

SLES 15 and FC-NVMe

SLES 15 was installed with all packages selected in the *Base System, Base development and Linux Kernel development* under the software section during the OS installation. Alternately, you may install the requisite packages using the zipper utility post the SLES15 OS install, after ensuring you have mounted the installer-1 and Packages-1 ISOs using the *Virtual Media Wizard* on your server.

Upgrade the SLES 15 OS image to Kernel 4.12.14-150.27.1-default version to avail the latest ANA fixes including the NVMe multipath load balancing fix.

```
#rpm -ivh kernel-default-4.12.14-150.27.1.x86_64.rpm kernel-default-devel-4.12.14-
150.27.1.x86_64.rpm kernel-macros-4.12.14-150.27.1.noarch.rpm kernel-devel-4.12.14-
150.27.1.noarch.rpm kernel-source-4.12.14-150.27.1.rpm
```

In SLES 15, NVMe multipath is enabled by default, and is the recommended NVMe in-kernel multipath config for enabling ANA.

You may verify this by running:

#cat /sys/module/nvme_core/parameters/multipath

Upgrade the server to the newer systemd-presets-branding-SLE-15.1.13.6.1 package and the nvme-cli-1.5-719.1 package. Please note this sequence- you need to first upgrade the systemd-presets-branding package before upgrading the nvme-cli:

#zypper install systemd-presets-branding-SLE-15.1-13.6.1.noarch.rpm

```
#zypper install nvme-cli-1.5-7.19.1.x86_64.rpm
```

ONTAP udev rule - New udev rule to ensure NVMe multipath round-robin loadbalancer default applies to all ONTAP namespaces:

```
# cat /usr/lib/udev/rules.d/71-nvme-iopolicy-netapp-ONTAP.rules
# Enable round-robin for NetApp ONTAP
ACTION=="add", SUBSYSTEM=="nvme-subsystem", ATTR{model}=="NetApp ONTAP Controller",
ATTR{iopolicy}="round-robin"
You can verify that the round-robin loadbalancer by checking the corresponding sysfs
```

value, provided of course you have ONTAP NVMe namespaces mapped and discovered on the host. For e.g.

```
# cat /sys/class/nvme-subsystem/nvme-subsys0/iopolicy
round-robin
```

NetApp plug-in for ONTAP devices - The existing NetApp plug-in has now been modified to handle ONTAP namespaces as well. It now provides 3 different display options for reporting NVMe specific ONTAP details - column, json & raw (which is the default) as shown in the below example:

```
# nvme netapp ontapdevices -o column
```

```
Device
                Vserver
                                         Namespace Path
   NSID UUID
                                             Size
_____
/dev/nvme0n1
               NVMEOF-svm01
                                         /vol/N NVME Fvol11/ns svm01 01
   1
       9e140f3a-4b08-4c4f-b962-487169343fe6
                                             16.11GB
/dev/nvme0n2
                NVMEOF-svm01
                                         /vol/N_NVME_Fvol12/ns_svm01_02
   2
       1148baad-3db0-438c-90ad-a3a08f0b5def
                                             16.11GB
#nvme netapp ontapdevices -o json
{
  "ONTAPdevices" : [
   {
     "Device" : "/dev/nvme0n1",
     "Vserver" : "NVMEOF-svm01",
     "Namespace Path" : "/vol/N NVME Fvol11/ns svm01 01",
     "NSID" : 1,
     "UUID" : "9e140f3a-4b08-4c4f-b962-487169343fe6",
     "Size" : "16.11GB",
     "LBA Data Size" : 4096,
     "Namespace_Size" : 3932160
},
{
     "Device" : "/dev/nvme0n2",
     "Vserver" : "NVMEOF-svm01",
```

#

The SLES 15 installer does not contain drivers for the Cisco VIC FNIC. Because of this, Fibre Channel SAN boot was not supported and SLES 15 had to be installed on local storage on the servers. In this lab validation, SLES 15 was installed on 32-GB Cisco FlexFlash mirrored Secure Digital (SD) drives on each server. Also, all the vHBAs (both Fibre Channel and FC-NVMe) had to be disabled in the service profile so that SLES could be installed. After SLES and the Release 2.0.0.42-77.0 fnic driver were installed, the vHBAs could be brought online. It was also discovered in this validation that the FC-NVMe vHBAs must come before the Fibre Channel vHBAs in the Cisco UCS service profile PCI device ordering.

The nvme command was used to connect to the NVMe subsystems provisioned on the NetApp storage.

# nvme lis	t			
Node	SN	Mode	1	Namespace
Usage		Format	FW Rev	
/dev/nvme0n1	815FbBMu2	LQnAAAAAAB NetA	pp ONTAP Controller	1
16.11 GB	/ 16.11 GB	4 KiB + 0	B FFFFFFF	
/dev/nvme0n2	815FbBMu2	LQnAAAAAAB NetA	pp ONTAP Controller	2
16.11 GB	/ 16.11 GB	4 KiB + 0	B FFFFFFF	
#				
#nvme list-su	bsys /dev/nvm	e0n1		
nvme-subsys0 -	- NQN=nqn.199	2–		
08.com.net	app:sn.6eb242	14818b11e9aae500	a098e217cb:subsystem.sle	es-01
١				
+- nvme0 fc +	traddr=nn-0x2	00c00a098e217ca:	pn-0x203000a098e217ca hc	ost_traddr=nn-
0x20000025	D2412000:bu-0	x2000002505a13a0	/ IIVE INACCESSIDIE	

- +- nvme1 fc traddr=nn-0x200c00a098e217ca:pn-0x203e00a098e217ca host_traddr=nn-0x20000025b5a13006:pn-0x20000025b5a13a07 live optimized
- +- nvme2 fc traddr=nn-0x200c00a098e217ca:pn-0x202900a098e217ca host_traddr=nn-0x20000025b5a13006:pn-0x20000025b5a13b07 live optimized
- +- nvme3 fc traddr=nn-0x200c00a098e217ca:pn-0x202d00a098e217ca host_traddr=nn-0x20000025b5a13006:pn-0x20000025b5a13b07 live inaccessible

#

In the above example, /dev/nvme0n1 refers to the NVMe multipath device as displayed in the *nvme list* output, and the underlying ANA path states are displayed in the *nvme list-subsys* /*dev/nvmeXnY* output.

The following capabilities were demonstrated in this validation:

- The ability to read data from and write data to a connected FC-NVMe device.
- The ability to simultaneously connect to both Fibre Channel and FC-NVMe targets.
- The ability to simultaneously connect to multiple FC-NVMe target devices using multiple FC-NVMe vHBAs.
- The ability to simultaneously connect to more than one FC-NVMe target device using a single FC-NVMe vHBA.

System validation and testing

This section provides a high-level summary of the FC-NVMe on FlexPod validation testing. The solution was validated for basic data forwarding by deploying the Vdbench tool on each SLES host. The system was verified to successfully pass FC-NVMe traffic, and performance comparisons were made between FC-NVMe and FCP with various data block sizes and numbers of Vdbench threads.

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:

- <u>NetApp Interoperability Matrix Tool</u>
- Cisco UCS Hardware and Software Interoperability Tool

Table 1. Validated hardware and software versions

Layer	Device	Image	Comments
Computing	Cisco UCS 6454 Fabric Interconnect with Cisco UCS B200 M5 Blade Server with VIC 1440 and Cisco UCS C220M5 Rack Server with VIC 1457	Release 4.0(4d)	
Network	Cisco Nexus 9336C-FX2 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 8.4(1)	Supports FC-NVMe San Analytics
Storage	NetApp AFF A800	NetApp ONTAP 9.6P2	
Software	SLES 15	Kernel 4.12.14-150.27-default	
	Cisco UCS VIC fnic driver	Release 2.0.0.42-77.0	Supports FC-NVMe

The FlexPod implementation used in this validation is based on <u>FlexPod Datacenter with VMware vSphere 6.7</u> <u>U1, Cisco UCS 4th Generation, and NetApp AFF A-Series</u> with modifications for SLES 15 with FC-NVMe.

Test results

Testing consisted of running the Vdbench tool to compare FCP performance and FC-NVMe performance. Because FC-NVMe is a more streamlined protocol with much greater queuing capabilities, it should outperform FCP with Fibre Channel and SCSI, especially in situations with more I/O operations per second (IOPS; that is, more transactions) and parallel activities.

The Test Setup Configuration Details:

```
Hosts:
 2 Cisco UCS B200 M5 servers
2 Cisco UCS C220M5 servers
Adapter:
Cisco UCS VIC1400 series adapter
Connections:
 FCP/FC-NVMe- 4 paths to each multipath device.
ONTAP Ports: 4 ports on each storage Node are enabled for FCP and FC-NVMe.
os:
SLES 15
      Modified the host Scheduler to deadline.
Modified the host sd queue depth to 32.
LUNs/Namespace:
16 LUNs each of size 15GB mapped to each Host server from the storage.
12 Namespaces each of size 20GB mapped to each Host server from the storage.
Test tool : Vdbench
```

For 4 KB Block size and 100% random reads, the following results were seen:



RndRead-4k Latency vs. IOPs

Figure 4.

4K Random Read Test Results

In this test,

- At least 30% lower latency is seen at ~680K IOPs
- ~65% increase in IOPs at ~250micro seconds. •

For 8 KB block size and 100% random reads, the following results were seen:



Figure 5.

8K Random Read Test Results

In this test,

- At least 24% lower latency is seen at ~600K IOPs.
- 57% increase in IOPs is seen at ~220 microseconds.

For a VDBench simulated Oracle workload at 80% Random Reads and 20% Writes the following results were seen:



Oracle-80-20

Figure 6.

VDBench Simulated Oracle Workload at 80% Random Reads and 20% writes Test Results

In this test,

- At least 25% lower latency is seen at ~700K IOPs •
- 56% increase in throughput is seen at ~300 microseconds. •

With respect to SLES 15 host CPU utilization, the following results were seen:



NVME vs FCP CPU utilization at 100% workload

Figure 7.

Host CPU Utilization Comparison

In all three cases, FC-NVMe utilized less CPU resources and less storage resources while providing better performance at lower latrency.

Conclusion

FlexPod Datacenter is the optimal shared infrastructure foundation for deploying FC-NVMe to allow highperformance storage access to applications that need it. As FC-NVMe evolves to include high availability, multipathing, and additional operating system support, 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 FC-NVMe, FlexPod adds another feature to help organizations efficiently and effectively support business-critical applications running simultaneously from the same shared infrastructure. The flexibility and scalability of FlexPod also 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 6454 Fabric Interconnect: <u>https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/datasheet-c78-</u> <u>741116.html</u>
- 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: <u>http://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c-series-rack-servers/index.html</u>
- Cisco UCS adapters:
 http://www.cisco.com/en/US/products/ps10277/prod_module_series_home.html
- Cisco UCS Manager:
 <u>http://www.cisco.com/en/US/products/ps10281/index.html</u>
- 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 A800: <u>http://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx</u>

Interoperability matrixes

- Cisco UCS Hardware Compatibility Matrix: <u>https://ucshcltool.cloudapps.cisco.com/public/</u>
- NetApp Interoperability Matrix Tool: <u>http://support.netapp.com/matrix/</u>
- Best Practices for modern SAN: <u>https://www.netapp.com/us/media/tr-4080.pdf</u>

Americas Headquarters Cisco Systems, Inc. San Jose, CA Asia Pacific Headquarters Cisco Systems (USA) Pte. Ltd. Singapore Europe Headquarters Cisco Systems International BV Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at https://www.cisco.com/go/offices.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: https://www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA