Introducing End-to-End NVMe for FlexPod

Last Updated: February 22, 2019
Contents

FlexPod overview .................................................................................................................................................. 3
  FlexPod design principles ................................................................................................................................. 4
  Latest FlexPod design ....................................................................................................................................... 4

NVMe storage and end-to-end FC-NVMe in FlexPod ......................................................................................... 7
  NetApp and FC-NVMe ....................................................................................................................................... 7
  Cisco UCS and FC-NVMe ................................................................................................................................. 8
  SLES 12 SP3 and FC-NVMe .............................................................................................................................. 9

System validation and testing ............................................................................................................................. 9
  Validated hardware and software ..................................................................................................................... 10
  Test results ....................................................................................................................................................... 10

Conclusion ............................................................................................................................................................ 10

For more information ......................................................................................................................................... 11
  Products and solutions ................................................................................................................................... 11
  Interoperability matrixes ................................................................................................................................. 11
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, 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 using IP.
This design 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.

Figure 3 shows the connection of the various elements of the latest FlexPod design in a Fibre Channel SAN.
This design also 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.

Note: In the Cisco UCS Hardware and Software Compatibility List the UCS C220 M5 Rack Server with Cisco UCS VIC 1457 is not currently listed as supported for the SLES kernel used with FC-NVMe and was not tested.

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.
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 9.4 is the first generally available enterprise storage that provides a complete NVMe over Fibre Channel, or FC-NVMe, solution. A nondisruptive ONTAP software update for AFF A300, 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:

- SUSE Linux Enterprise Server (SLES) 12 SP3 is the only enterprise host OS supported by ONTAP 9.4. ONTAP 9.5 also supports SLES 15 and Red Hat Enterprise Linux (RHEL) 7.6.
FC-NVMe is currently supported only on 32-Gbps end-to-end configurations of ONTAP 9.4.

ONTAP 9.4 assigns a single logical interface (LIF) for each FC-NVMe storage virtual machine (SVM). However, this limitation is lifted with ONTAP 9.5. ONTAP 9.5 supports multiple LIFs with the enterprise host OS SLES 15.

FC-NVMe path failover is not supported in ONTAP 9.4.

**Note:** Asymmetric namespace access (ANA), which formalizes NVMe multipathing and path failover support, was ratified in March 2018 and will be adopted by host OS vendors. It is added to ONTAP 9.5.

Path failover is supported with ONTAP 9.5 and SLES 15.

- The SLES 12 SP3 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 `dm-multipath` as usual for the SCSI logical unit numbers (LUNs), or logical devices, resulting in multipath devices. This configuration allows normal multipath operation for those FCP LUNs. However, it does not affect the FC-NVMe namespace devices (for example, `/dev/nvmeXnY`) on the SLES 12 SP3 host, which continue to be single pathed and do not support path failover.
- SVMs (referred to as Vservers in the command-line interface [CLI]) can support either FC-NVMe or all the other ONTAP block and file protocols. This restriction was put in place for the ONTAP 9.4 rollout to accelerate and simplify development. It also reduces the possibility that the FC-NVMe target code may introduce a regression in any of the other ONTAP protocols. This restriction is removed in the ONTAP 9.5 release.
- The NVMe service must be started before a LIF is created.
- NVMe LIFs and namespaces must be hosted on the same node.
- Only one NVMe LIF is permitted per SVM. This restriction will be removed in an upcoming version of ONTAP.
- LUNs and namespaces cannot be mixed on the same volume. This restriction will be removed in an upcoming version of ONTAP.

**Note:** FC-NVMe is under rapid development, versions can change, and the number of items qualified continue to expand rapidly. For this reason, and because it is a best practice, always check the Interoperability Matrix Tool (IMT) and adjust versions and hardware specifications to those listed in the IMT.

**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(2), along with the Cisco UCS VIC 1400 platform adapters and the Release 2.0.0.27 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 4.
**Figure 4.** FC-NVMe vHBA type selection

<table>
<thead>
<tr>
<th>vHBA Type</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC Initiator</td>
<td>✓</td>
</tr>
<tr>
<td>FC Target</td>
<td>✓</td>
</tr>
<tr>
<td>FC NVMe Initiator</td>
<td>✓</td>
</tr>
<tr>
<td>FC NVMe Target</td>
<td></td>
</tr>
</tbody>
</table>

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 12 SP3 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 NPIV mode. Zoning in the 9132T switches connected the vHBAs to storage targets for both FC-NVMe and Fibre Channel. Because of the current single-path limitation with FC-NVMe, single-initiator, single-target zones were used for FC-NVMe. The connectivity to the NVMe subsystems on storage is configured in both the storage and the host operating system.

**SLES 12 SP3 and FC-NVMe**

The SLES 12 SP3 installer does not contain drivers for the Cisco VIC fnic. Because of this, Fibre Channel SAN boot was not supported and SLES 12 SP3 had to be installed on local storage on the servers. In this lab validation, SLES 12 SP3 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 2.0.0.27 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.

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 `fio` 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 `fio` processes. In order to create an accurate comparison between FCP and FC-NVME, for FCP a single path to storage was configured and multipathing was disabled.
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

<table>
<thead>
<tr>
<th>Layer</th>
<th>Device</th>
<th>Image</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computing</td>
<td>Cisco UCS 6454 Fabric Interconnect and Cisco UCS B200 M5 Blade Server with VIC 1440</td>
<td>Release 4.0(2a)</td>
<td>The Cisco UCS C220 M5 Rack Server is not currently listed as supported in the Cisco UCS HCL and was not tested.</td>
</tr>
<tr>
<td>Network</td>
<td>Cisco Nexus 9336C-FX2 Switch in NX-OS standalone mode</td>
<td>Release 9.2(2)</td>
<td></td>
</tr>
<tr>
<td>Storage network</td>
<td>Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch</td>
<td>Release 8.3(2)</td>
<td>Supports FC-NVMe</td>
</tr>
<tr>
<td>Storage</td>
<td>NetApp AFF A800</td>
<td>NetApp ONTAP 9.4P3</td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td>SLES 12 SP3</td>
<td>Kernel 4.4.126_94.22-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cisco UCS VIC fnic driver</td>
<td>Release 2.0.0.27</td>
<td>Supports FC-NVMe</td>
</tr>
</tbody>
</table>

Test results

Testing consisted of running the fio 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 for this solution used two Cisco UCS B200 M5 worker servers with Cisco UCS VIC 1440 adapters running SLES 12 SP3 and fio. Each server had one FC-NVMe and one FCP mounted device that was partitioned and formatted with XFS. Both devices were connected on a single Fibre Channel path from storage.

A single fio test file was placed on each mounted disk. fio tests were run with one process and with sixteen processes with 4-, 8-, 16-, 32-, and 64-KB block sizes with 100 percent read and 100 percent random disk access. In tests with one process, only a small performance difference was seen. In tests with sixteen processes, in every case except the 64-KB block size, FC-NVMe outperformed Fibre channel. Larger performance differences were seen as block sizes decreased (higher IOPS count).

Conclusion

FlexPod Datacenter is the optimal shared infrastructure foundation for deploying FC-NVMe to allow high-performance storage access to applications that need it. As FC-NVMe evolves to include high availability, multipathing, and additional server 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

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