NetApp Verified Architecture

ML and AI on NetApp EF-Series and Brocade Fibre Channel fabrics with Spectrum Scale

NVA design

Tim Chau, NetApp
Naem Saafein, PhD, Brocade
November 2022 | NVA-1169-DESIGN | Version 1.0

Abstract

This document describes a NetApp Verified Architecture for artificial intelligence (AI) workloads using NetApp® EF600 NVMe storage systems, Spectrum Scale file system, Brocade Gen 7 Fibre Channel switches, and Emulex Gen 7 HBAs. This design features 64Gbps FC for the storage and compute cluster interconnect fabric to provide customers with a completely SAN-based architecture for high-performance workloads. This document also includes benchmark test results for the architecture as implemented.

In partnership with

BROADCOM
EMULEX | BROCADE
FIBRE CHANNEL SOLUTIONS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Executive summary

As most enterprises are currently striving to leverage AI technologies as an enabler for new services, the road of deep learning (DL) is facing many challenges, particularly in the areas of data management, performance expectations, and the handling of mega datasets. An AI infrastructure should not only enable DL but meet the growing demands of AI workflow in terms of I/O, latency, and performance, where massive I/O bandwidth with extreme I/O parallelism is needed to feed data to the DL training cluster for processing, followed by accessing that data that is often stored in a DevOps-style repository where ultra-low-latency access is a must.

NetApp offerings such as NetApp E-Series, NetApp AFF, Data Fabrics, NetApp ONTAP® data management software coupled with a trusted and proven SAN Networks from Brocade will address storage needs both high performance and high capacity while reducing the need for time-consuming data copies. NetApp and Brocade are delivering NVMe over Fibre Channel Fabrics (FC-NVMe) to further extend E-/EF-Systems capabilities. FC-NVMe is a specification that extends the benefits of NVMe to larger fabrics beyond the reach and scalability of a single server. It enables NVMe message-based commands to transfer data between a host computer and a target NVMe SSD over a high-performance FC network. An end-to-end NVMe solution reduces access latency and improves performance, particularly when paired with a low-latency, high-efficiency transport such as FC.

The Brocade Gen 7 coupled with Fabric Vision technology includes VM Insight and IO Insight to help organizations not only achieve greater visibility into performance monitoring, but also help ensure that critical SLAs are met by monitoring I/O statistics, including device latency and IOPS metrics, to provide intelligence for early detection of storage performance degradations. FC remains a core part of today's IT infrastructure because of its performance, reliability, scalability, and availability and is expected to play a vital role in the future of advanced storage deployments.

NetApp, in partnership with Brocade, is enabling its customers to deliver superior IT performance for their most important, mission-critical enterprise SAN applications. This document provides details on how to design an IBM Spectrum Scale file system SAN mode solution. NetApp E-Series storage systems, Emulex Gen 7 LPe35000/36000 series FC host bus adapters (HBAs), and Brocade Gen 7 SAN fabric switches are used to deliver this solution, which is ideally suited for the AI workloads in an FC environment. This solution outlines the all-NVMe NetApp EF600 all-flash array and offers performance characterization based on the FIO benchmarking tool and MLperf which are both used in the high-performance computing (HPC) and AI industry for testing.

This document contains validation information for the NetApp EF-Series AI reference architecture for machine learning (ML) and artificial intelligence (AI) workloads. This design was implemented using a NetApp EF600 all-flash NVMe storage system, the Spectrum Scale parallel file system, one Supermicro AS-2124GQ-NART-LLC systems, and G720 switches for both the compute cluster interconnect and storage connectivity. The operation and performance of this system was validated using industry-standard benchmark tools and proven to deliver excellent training performance. Customers can easily and independently scale compute and storage resources from half-rack to multi-rack configurations with predictable performance to meet any machine learning workload requirement.

Use case summary

As HPC and AI workloads in businesses become more common place, the need to deploy infrastructure to support these new workloads is a hard requirement. Expanding and upgrading the existing high-performance, easy-to-manage, highly available, and high-performing FC SAN fabrics is the ideal choice for system administrations and other IT professionals in supporting the new demanding HPC big data and AI workflows.

This solution outlined below applies to the following use cases that require a high-performance network within a Spectrum Scale environment:

- The ingest of large dataset from edged devices
- Data preprocessing to normalize and cleanse data before AI training
- Training phase in DL pipeline
- Big data analysis on large dataset

Solution overview

With leading-edge NVMe technology, NetApp E-Series, together with Brocade Gen 7 Fibre Channel switches, Emulex Gen 7 HBAs, and a parallel file system such as Spectrum Scale, dramatically streamlines workflow and improves productivity. This combination creates a shared repository that supports:

- Flexibility
- Reliability
- Unmatched predictability
- High-performance streaming
- Massive scalability

This shared repository also includes:

- A single namespace and limitless bandwidth or capacity
- Near-linear bandwidth scalability; both scale-up and scale-out configurations
- Supports direct access for Linux and Windows clients while macOS access is provided through SMB

Businesses deploying HPC and AI workloads are challenged to find storage tier solutions that both satisfy their high-density, high-bandwidth requirements, while also optimizing rack space power and cooling. The NetApp EF600 array fulfills these solution requirements.

The architecture demonstrated in this design guide shows the capabilities of a single, high-performance building block by using a single EF600 array with high-performing NVMe/FC drives. Additional EF600 arrays can be added to the IBM Spectrum Scale cluster to enable unlimited scale out of performance and capacity.

The value of this solution comes from the proven ability of NetApp storage architecture and Brocade FC high-speed network fabrics to deliver high-throughput performance while maintaining a low-latency profile. This solution comes in a 3U rack space, which can provide savings in both footprint, power and cooling costs.

Solution technology

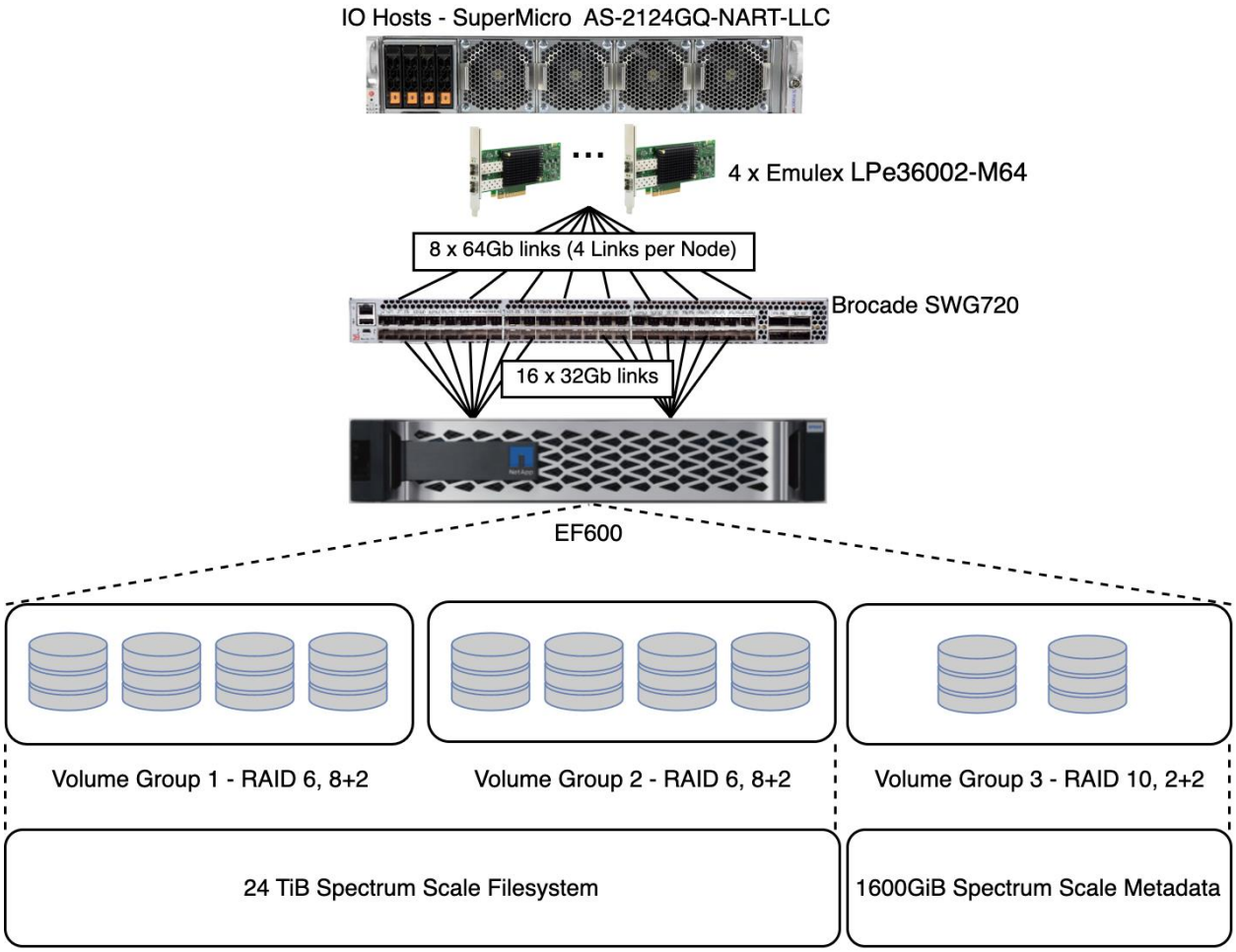
Using a proven validated design by NetApp and Broadcom's Brocade and Emulex divisions, NetApp and Broadcom provide an end-to-end NVMe-powered FC solution, from host to storage controller. This solution can help you realize the promise and benefits of NVMe/FC technology. With a system that yields the fastest access, simplified management, and utilization of critical data, you can leverage the following features:

- This Spectrum Scale solution consists of one Spectrum Scale client with 4x A100 40GB GPUs that read and write to an EF600 array.
- This solution also includes SAN fabric components from Broadcom's Brocade industry leading 64Gb FC switches and Emulex Gen7 HBAs.
- The EF600 array has 24x NVMe/FC drives and is provisioned into two 8+2 RAID 6 volume groups. These groups are striped into a single Spectrum scale file system and one 2+2 RAID 10 volume group used for metadata storage.

- With NVMe/FC support, Broadcom Gen 7 FC Fabrics, and NetApp E-series arrays, we are ready to enter a new era of wider adoption of NVMe for HPC and AI environments. This is possible today with NetApp storage and the industry's first end-to-end NVMe/FC platform.

Figure 1 shows the technical components of the EF600 solution.

Figure 1) EF600 technical overview.



Technology requirements

This section covers the technology requirements for the Spectrum Scale Brocade and NetApp SAN solution.

Hardware requirements

This section provides the hardware components that are required to implement the solution.

Server: SuperMicro client

- SuperMicro AS-2124GQ-NART-LLC:
  - CPU - 2x AMD EPYC 7742 64-Core Processor
  - Memory – 1TB
  - GPU – 4 x NVIDIA HGX A100 GPU 40GB

Storage network: Broadcom FC SAN Fabric

- FC SAN fabric:
  - 4x Emulex LPe36002-M64 2 Port Gen 7 64G HBAs
  - 1x Brocade G720 Gen7 64G FC Switch

Storage: NetApp E-Series arrays

- EF600 with 16x 32Gb FC ports:
  - 24x MZWLJ1T9HBJR-0G3– 1.6TB NVMe drives
  - 2x 8+2 RAID 6 volume groups (data volumes)
    - 4x 2.92TiB volumes per volume group with 512KiB segment size (capacity is 16% under provisioned per NVMe drive guidelines)
  - 1x 4 drive RAID 10 volume group (metadata volumes)
  - 4x 400GiB Volumes with 128KiB segment size

Software requirements

This design includes Spectrum Scale file system version 5.1.4-1. Spectrum Scale file system is a versatile, high-performance shared file system that offers heterogeneous, block-based access to a single storage system or striped across multiple external storage systems.

This example consists of tests that were performed to an EF600 storage target running NetApp SANtricity® firmware version 11.73.

Although the clients in this design were running Red Hat Enterprise Linux (RHEL) 8.4, with Native NVMe failover support, a wide variety of host operating systems are supported. For a full listing of support operating system combinations, consult the NetApp Interoperability Matrix.

Table 1 lists the software components that are required to implement the solution. The software components that are used in any particular implementation of the solution might vary based on customer requirements.

Table 1) Software requirements.

Software	Version or other information
Client OS	RHEL 8.4
Spectrum Scale	5.1.4-1
Brocade FOS	v9.0.1a
Emulex lpfc driver/FW	Driver: 12.8.0.5 / FW: 12.8.351.37
SANtricity	11.73

Solution verification

For this solution, NetApp evaluated the performance of EF600 storage systems with Broadcom's Brocade and Emulex divisions SAN Fabric to determine the performance of using NVMe/FC to support AI workloads using synthetic benchmark utilities and DL benchmark tests to establish baseline performance and operation of the system. Each of the tests described in this section was performed with the specific equipment and software listed in "Technology requirements."

FIO bandwidth and IOPS tests

These tests are intended to measure the storage system performance using the synthetic I/O generator tool FIO. Two separate configurations were used, one optimized to deliver maximum bandwidth and the

other optimized for IOPS. Each configuration was run with both 100% reads and 100% writes, and the files used by FIO were created as a separate step to isolate those activities from the actual test results.

Here are the specific FIO configuration parameters for these tests:

- ioengine = posixaio
- direct = 1
- blocksize = 1MB for bandwidth test, 4k for IOPS test
- numjobs = 1 to 512
- iodepth = 1 for Bandwidth Test and 16 for IOPS Test
- File size per job = 8g

Test results

Figure 2 shows the bandwidth results with the number of Jobs scaled from 1 to 512 on the Supermicro client. Write bandwidth performance scales up to 32 jobs before peaking at 24GiBps due to limits in the EF600. Read bandwidth performance scales up to 512 jobs with a peak read bandwidth of 40GiBps.

Figure 2) FIO bandwidth results - Single client process scaling.

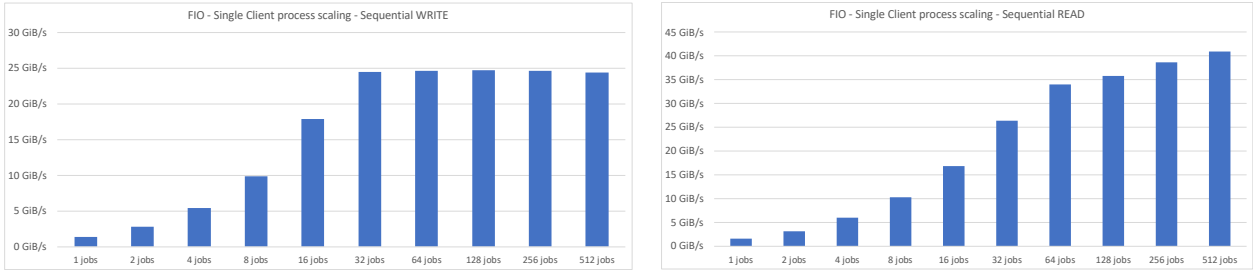
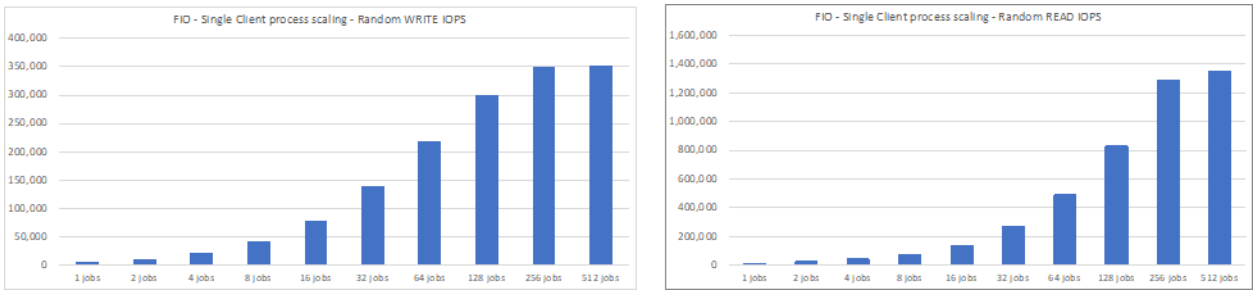


Figure 3 shows the IOPS results with the number of jobs scaled from 1 to 512 on the HGX client. Write IOPS performance scales up to 256 jobs before peaking at 350 K IOPS while read IOPS performance scales up to 512 jobs with a peak read IOPS of 1.3 million IOPS.

Figure 3) FIO IOPS results - Single client process scaling.

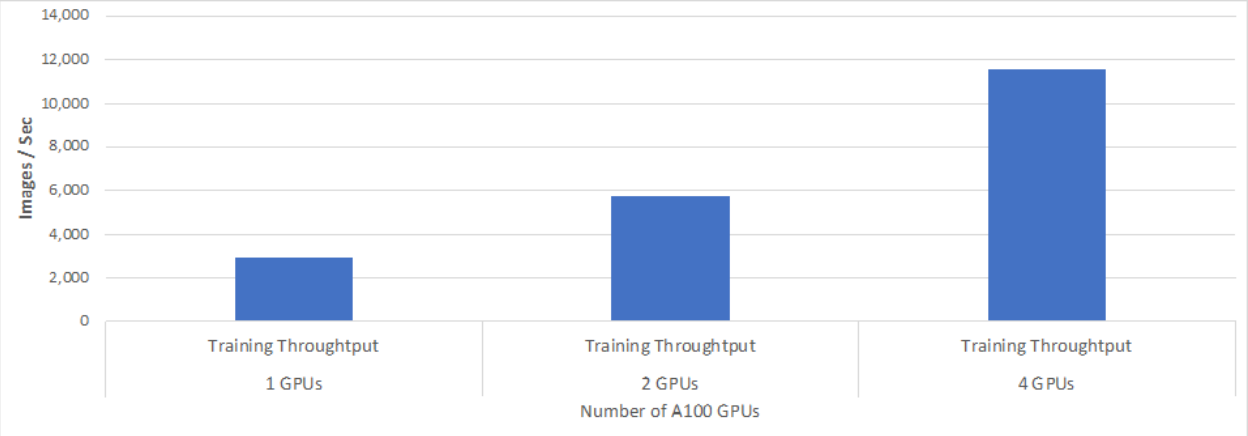


MLPerf Training v1.1 ResNet-50

This reference architecture was tested using a MLPerf Training v1.1 benchmark to validate the operation of DL workloads on the deployed infrastructure. MLPerf is an industry-standard benchmark implementation of various neural networks for validating the performance of DL infrastructure. This test used the MXNet implementation of ResNet-50 in addition to the ImageNet dataset in IORecord format to validate model training performance. The results presented maintain a consistent batch size per system of 204 images as the workload is scaled GPUs.

The base container image used for these tests is the 21.09 MXNet image from NGC. MLPerf benchmark tests are deliberately not optimized for any specific hardware implementation so overall system performance in these tests can be increased by tuning parameters such as concurrency.

Figure 4) MLPerf Training v1.1 ResNet-50 average images per second over 8 epochs.



Note: These are unverified scores of v1.1 on the MLPerf image classification benchmark and have not been submitted for review or validation.

E-Series EF600

All-flash EF600, NetApp's first all-NVMe EF-Series platform, can accelerate access to your data so you can derive value from it faster. The EF600 doubles the performance of SAS all-flash arrays. You can accelerate write IOPS and read/write throughput with an end-to-end NVMe storage platform that's purpose-built for high-performance workloads. It offers the most powerful performance, smart value, and trusted simplicity in dense, 2U enterprise packaging to derive faster, more actionable results. It also unlocks the value in your data and rapidly develop insights that were previously unrealistic for performance-sensitive AI workloads, real-time analytics, and high-performance computing applications on top of a Spectrum Scale high-performance parallel file system. The EF600 all-flash array combines extreme IOPS, response times of less than 100 microseconds, and up to 44GBps of bandwidth with leading, enterprise-proven availability features. The NetApp EF-Series solution for high-performance Oracle databases enables you to leverage best-in-class, end-to-end, modern SAN and NVMe technologies to deliver business-critical IT services today while preparing for the future. With the EF-Series, NetApp has created a SAN array that is both future-ready and usable today—and it's easy to implement with your current operational processes and procedures.

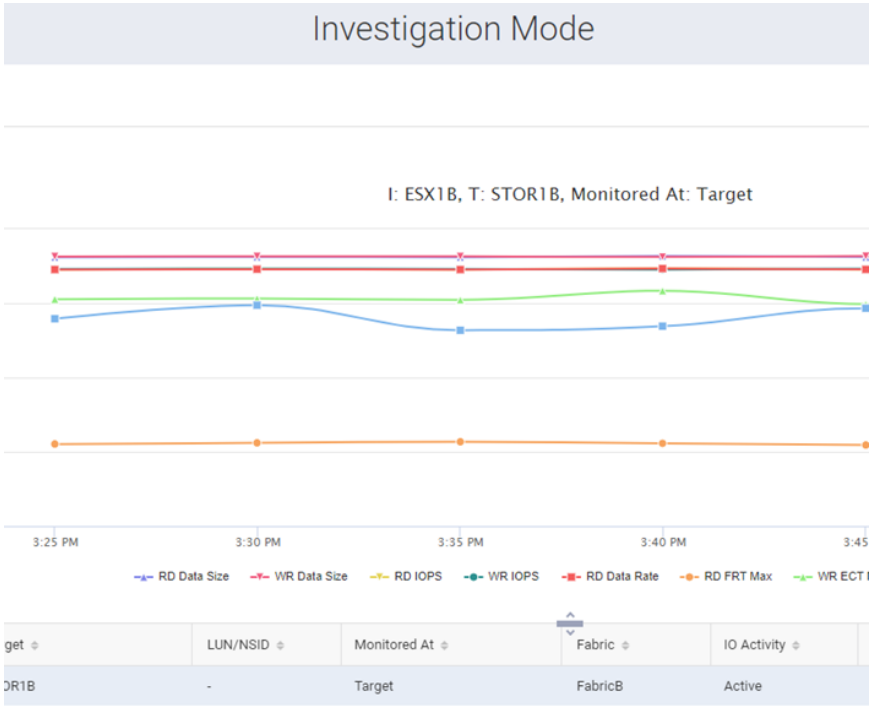
Brocade Gen 7 Fibre Channel Platforms

Broadcom's Brocade has been the leading provider of storage networking solutions worldwide for more than 20 years, supporting the mission-critical systems and business-critical applications of most large enterprises Brocade supports the high-end network infrastructure requirements and demands of today's AI workloads. As datasets for AI continue to grow, Brocade offers industry-leading network reliability, scalability, and security to support tomorrow's most demanding workloads.

Brocade's expertise in high-performance switching and NetApp's latest generation of E-Series storage arrays and management software expertise combine to deliver compelling AI and HPC solution sets. Brocade FC SAN fabrics provide industry-leading high availability and reliability that enable organizations to reliably run HPC and AI workloads with a maximum throughput across all network connections in the absence of a failure.

A Brocade Gen 7 Fibre Channel infrastructure also unleashes the performance of NVMe workloads with reduced latency and increased bandwidth. In addition, Brocade fabrics offer integrated intelligence and automation that are built on analytics and telemetry data to further simplify and optimize the environment. This infrastructure lays the foundation for an autonomous SAN by combining powerful analytics and advanced automation capabilities to maximize performance and ensure reliability. Also, Brocade Fabric Vision technology is a suite of features that leverage comprehensive data collection capabilities with powerful analytics to quickly understand the health and performance of the environment and identify any potential impacts or trending problems. While VM Insight seamlessly monitors virtual machine (VM) performance throughout a storage fabric with standards-based, end-to-end VM tagging to quickly determine the source of VM and application performance anomalies to provision and fine-tune the infrastructure, Brocade products also proactively monitor I/O performance and behavior data points through integrated network sensors to gain monitor I/O to gain deep insight into the environment, as shown in Figure 5.

Figure 5) I/O Insight metrics displayed in Brocade's SANnav real-time Investigation Mode.



Emulex Gen 7 FC HBAs

Emulex FC HBAs by Broadcom are designed to meet the demanding performance, reliability, and management requirements of modern networked storage systems that use high-performance and low-latency SSDs. The Emulex Gen 7 LPe35000/36000 series FC HBAs with Dynamic Multi-core Architecture delivers unparalleled performance and more efficient port usage than other HBAs by applying all ASIC resources to any port that needs it, providing industry-leading 32Gb FC performance of over 5 million IOPs and over 11 million IOPS for 64Gb FC. The LPe35000/36000 series delivers 12800MBps (two 32Gb FC ports) full duplex, and three times better hardware latency than previous generation adapters. Emulex Gen 7 HBAs running NVMe/FC deliver extreme low latency- up to 55% lower insertion latency for NVMe/FC than SCSI over FC. With the ability to run both NVMe/FC and SCSI FCP concurrently, Emulex provides investment protection by enabling data centers to transition to end-to-end NVMe over FC SANs at their own pace. The secure firmware update feature protects and ensures the authenticity of device firmware.

Emulex works closely with its enterprise customers, developing tools aimed at lowering the cost of management. Emulex SAN Manager is a free, easy-to-use solution that dramatically reduces the operational cost and complexity of running a FC SAN.

Conclusion

Data storage requirements in HPC and AI have increased drastically due to the increased need of larger datasets for big data and DL applications. NetApp and Brocade have partnered together to offer a solution that addresses these unique challenges and requirements.

Using Spectrum Scale with the NetApp EF600 and Brocade SAN Gen 7 family provides a high-performing, shared storage system. It is optimized to support data and performance requirements for HPC and AI workloads, in only 3U of rack space.

With the EF-Series, NetApp, in collaboration with Brocade and Emulex, has created a SAN array that is both future-ready and usable today. It's easy to implement with your current operational processes and procedures. The NetApp EF-Series flash array plus Broadcom's Emulex and Brocade 32Gb and 64Gb FC HBAs and switches are market leaders in delivering high performance, consistent low latency, and advanced HA features for your NVMe/FC environment.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NetApp Artificial Intelligence Solutions
  https://www.netapp.com/artificial-intelligence/
- NetApp HPC solution:
  https://www.netapp.com/artificial-intelligence/high-performance-computing
- Leading the Future of Flash with NVMe
  www.netapp.com/us/info/nvme.aspx
- Brocade Fibre Channel networking products
  https://www.broadcom.com/products/fibre-channel-networking/directors/x7-directors
  https://www.broadcom.com/products/fibre-channel-networking/switches/
- Brocade and NetApp partner documents
  https://www.broadcom.com/company/oem-partners/fibre-channel-networking/netapp
- Emulex Gen 7 HBA
  https://www.broadcom.com/products/storage/fibre-channel-host-bus-adapters/lpe36002-m64

Version history

Version	Date	Document version history
Version 1.0	November 2022	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NVA-1169-DESIGN-1122

NetApp