NetApp Verified Architecture

Spectrum Scale SAN mode with NetApp EF600 and Brocade FC fabrics for HPC/AI

NVA design

Tim Chau, NetApp
Naem Saafein, PhD, Brocade
January 2022 | NVA-1162-DESIGN | Version 1.0

Abstract

NetApp and Broadcom (Brocade) have been OEM partners for more than two decades providing FC SAN fabric solutions to customers for more than two decades. NetApp continues to maintain its market leadership position in Non-Volatile Memory Express (NVMe) over Fibre Channel (NVMe/FC) technology. 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 parallel file system SAN mode solution. NetApp® E-Series storage systems, Emulex host bus adapters (HBAs), and Brocade SAN fabric switches are used to deliver this solution, which is ideally suited for the high-performance computing (HPC)/AI workloads in an FC environment. This solution outlines the all-NVMe NetApp EF600 all-flash array and offers performance characterization based on the IO500 benchmarking suite, which contains Interleaved or Random (IOR) and mdtest, which are both used in the HPC/AI industry for testing.

In partnership with

TABLE OF CONTENTS

Executive summary **3**

Use case summary **3**

Solution overview **3**

Solution technology 4

Technology requirements **5**

Hardware requirements 5

Software requirements 6

Solution verification **6**

Test results 6

Conclusion **8**

Where to find additional information **8**

Version history **8**

LIST OF TABLES

Table 1) Software requirements. 6

LIST OF FIGURES

Figure 1) EF600 technical overview. 5

Figure 2) IOR - Easy single client process scaling. 7

Figure 3) IOR – Easy multiclient Scaling. 7

Figure 4) IO500 results 7

Executive summary

Brocade supports the high-end network infrastructure requirements and demands of today's HPC and AI workloads. As datasets for HPC and 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 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.

The Brocade FC SAN fabric features include:

- Reliability and throughput performance for applications that require load balancing and high availability (HA)
- Fabric OS (FOS) advanced protocols that efficiently transport large datasets with the correct priority for the application
- Nondisruptive upgrade; very strong resiliency
- Power and cooling efficiencies; performance in high-density environments
- High-performance networks that ensure ability to scale and provide fast access to HPC workloads as required
- Network resiliency and redundancy
- Ultra-low latency
- End-to-end full native support for NVMe over FC

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 deep learning pipeline
- Big data analysis on large dataset

Solution overview

With leading-edge NVMe technology, NetApp E-Series, together with Brocade Gen 7 FC switches, Emulex HBAs, and a 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.

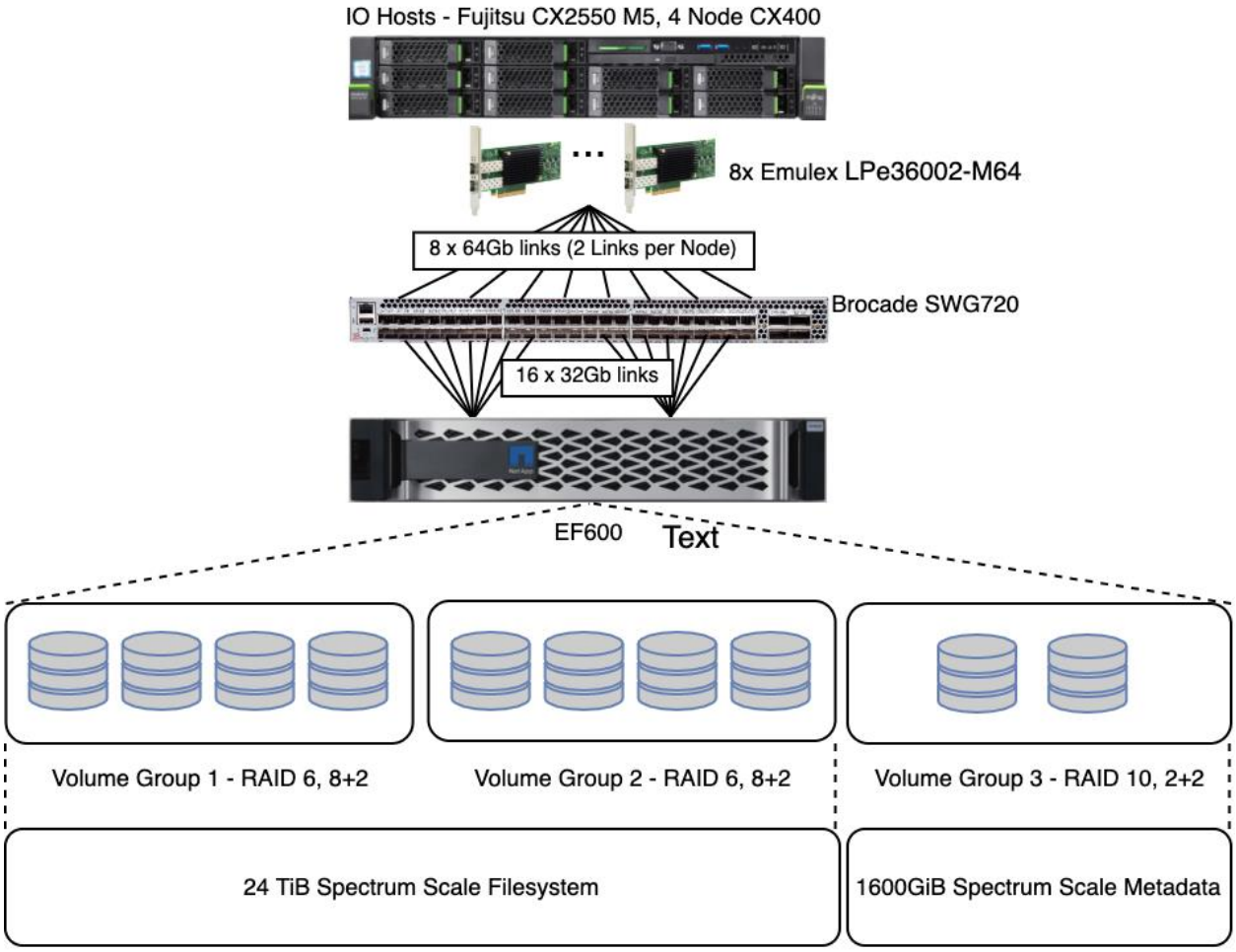
Solution technology

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 at least four Spectrum Scale clients 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: Fujitsu multinode server

- Fujitsu CX400 M4 w/ 4x CX2550 M5 nodes:
  - Per CX2550 M5 Node CPU - 2x Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz
  - Per CX2550 M5 Node Memory - 192GB

Storage network: Broadcom FC SAN Fabric

- FC SAN fabric:
  - 8x Emulex LPe36002-M Gen6 HBA (2x per CX2550 M5 node)
  - 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 5.0.5-1. 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 11.70.

Although the clients in this design were running Red Hat Enterprise Linux (RHEL) 7.8, with Device Mapper for 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 7.8
Spectrum Scale	5.0.5-1
Brocade FOS	v9.0.1a
Emulex lpfc driver/FW	Driver: 12.8.351.29 / FW: 12.8.351.37
SANtricity	11.70

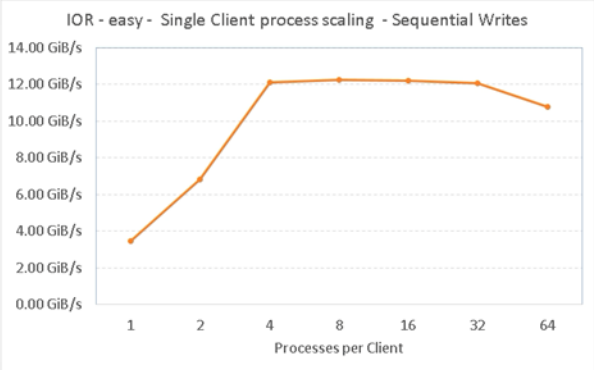
Solution verification

For this solution, NetApp studied the performance of EF600 storage systems with Broadcom's Brocade and Emulex divisions SAN Fabric to determine the performance gain of using NVMe/FC to support HPC and AI workloads. The IO500 benchmarking suite (https://www.vi4io.org/io500/about/start) contains twelve benchmarks that measure large sequential I/O, small random I/O, various metadata operations, as well as a find test. The file IO500 score is calculated as the geometric mean of the results of all 12 tests, and can be useful to help evaluate a parallel file system.

Test results

Figure 2 shows the bandwidth results as the number of processes are scaled from 1 to 64 for a single client. Note that four 64Gb FC links were available on the clients, but only two 64Gb links were used, one on each adapter due to the bandwidth limits of the x8 PCIe 3.0 slots of the clients. With that said, we are able to fully saturate the two 64Gb FC links on the clients with a peak bandwidth of 12GBps for both reads and writes.

Figure 2) IOR - Easy single client process scaling.



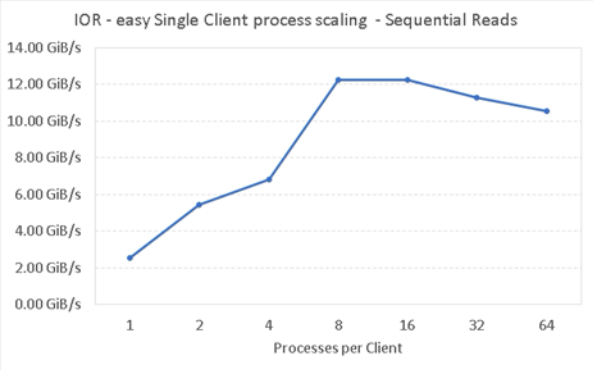
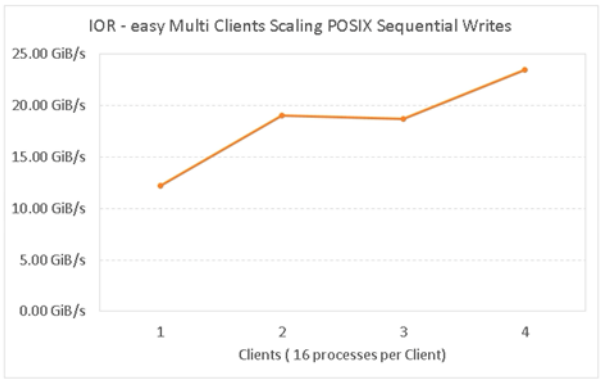
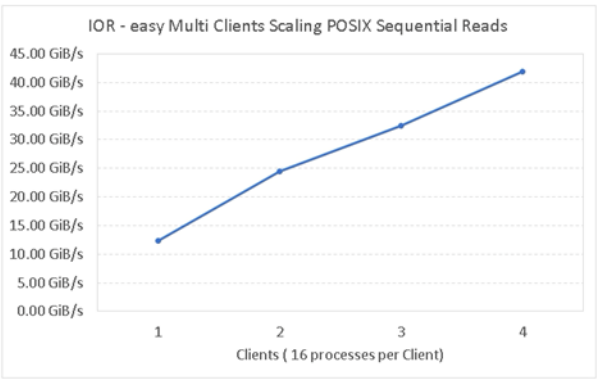


Figure 3 shows the bandwidth results as we scale the number of clients from one to four, with 16 processes per client. Read performance scales linearly as we add clients up to 41GiBps reads while writes scale linearly up to two clients with a peak bandwidth of 23GiBps with four clients which are close to the hardware limits of the EF600.

Figure 3) IOR – Easy multiclient Scaling.





The IO500 performance results of this solution are listed below. These results were achieved with four client nodes running 16 tasks per client for a total of 64 tasks.. For the IOR-easy-write test, the bandwidth achieved was 23GiBps while the IOR-easy-read test achieved 41GiBps, as highlighted in green.

Figure 4) IO500 results .

```
IO500 version
[RESULT]       ior-easy-write        23.437125 GiB/s : time 301.401 seconds
[RESULT]    mdtest-easy-write        80.447231 kIOPS : time 361.318 seconds
[RESULT]       ior-hard-write         1.764726 GiB/s : time 300.385 seconds
[RESULT]    mdtest-hard-write         5.368230 kIOPS : time 438.826 seconds
[RESULT]                 find       594.716741 kIOPS : time 51.406 seconds
[RESULT]        ior-easy-read        41.921770 GiB/s : time 168.496 seconds
[RESULT]     mdtest-easy-stat       149.558669 kIOPS : time 189.003 seconds
[RESULT]        ior-hard-read         3.182525 GiB/s : time 166.554 seconds
[RESULT]     mdtest-hard-stat        16.950686 kIOPS : time 135.231 seconds
[RESULT]   mdtest-easy-delete        49.087391 kIOPS : time 579.176 seconds
[RESULT]     mdtest-hard-read         5.082568 kIOPS : time 450.944 seconds
[RESULT]   mdtest-hard-delete         2.721328 kIOPS : time 842.291 seconds
[SCORE] Bandwidth 8.618829 GiB/s : IOPS 28.555230 kiops : TOTAL 15.687978
```

It should be noted that IO500 runs only a single iteration for each of the 12 tests, so run-to-run variations on the same clients and solution are to be expected, particularly in the “hard” test cases. The results

displayed here should not be seen as a performance guarantee, but only as a guideline to demonstrate the potential performance capabilities of this solution. Your mileage might vary.

Also note that number client nodes and the number of tasks per node heavily affect the IO500 results. It is recommended to run the tests with the number of clients and/or task per node that best represents the expected workload.

Conclusion

Data storage requirements in HPC/AI have increased drastically due to the increased need of larger dataset for big data and deep learning 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/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
- NetApp Product Documentation
  https://docs.netapp.com
- 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	January 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.

Copyright Information

Trademark Information

NVA-1162-DESIGN-0122