NetApp Verified Architecture

BeeGFS on NetApp EF-Series with NVIDIA DGX SuperPOD Systems

NVA Deployment

January 2025 | NVA-1174-DEPLOY

Abstract

This NetApp Verified Architecture describes the design of the NVIDIA DGX SuperPOD™ with NetApp® BeeGFS® building blocks. This solution is a full-stack data center platform that is validated on a dedicated acceptance cluster at NVIDIA.

TABLE OF CONTENTS

Executive summary **5**

Solution overview **5**

Use case summary 5

High Availability Architecture 7

Technology requirements **7**

NetApp BeeGFS hardware requirements 7

Software requirements 8

User storage 9

Data movement 9

Power requirements 9

Hardware installation and basic setup 9

Sizing guidelines **9**

Performance sizing 10

Capacity sizing for converged building blocks 11

Capacity sizing for storage-only building blocks **11**

DGX SuperPOD network fabrics **11**

NetApp BeeGFS building block connections **12**

In-band management network 14

Out-of-band management network 15

Designs and Architectures **17**

Single DGX SuperPOD SU with three building blocks 17

Scaling DGX SuperPOD with building blocks 18

Deployment procedures **19**

DGX SuperPOD deployment and configuration using Ansible 19

Ansible control node setup 19
Hardware deployment 19
Software deployment 21
BeeGFS client installation and configuration 21
Optimizing performance with BeeGFS 22
Solution verification 22
FIO storage throughput benchmark test 22
FIO storage IOPS benchmark test 22
MLPerf Training v0.7 benchmark test 23
Conclusion 23
Bill of materials (BOM) for 1 SU 23
Where to find additional information 25
NetApp EF-Series systems 25
NetApp Interoperability Matrix Tool 25
NVIDIA DGX systems 25
NVIDIA Networking 25
Machine learning frameworks 25

LIST OF FIGURES

Figure 1) A high-level view of the solution. 6
Figure 3) DGX SuperPOD storage network architecture 12
Figure 4) Three building blocks cabled to the DGX SuperPOD storage fabric 13
Figure 5) Lenovo SR665 V3 to network cabling 133
Figure 6) Lenovo SR665 V3 to EF600 cabling 14
Figure 7) Building block OOB management 15
Figure 8) File node NUMA configurations. 16
Figure 9) Single DGX SuperPOD with three building blocks storage architecture 17
Figure 10) Scaling DGX SuperPOD with EF600 building blocks storage architecture 18
Figure 11) Lenovo SR665 V3 host card adapter layout 20
Figure 12) Lenovo SR665 V3 host card adapter layout 20

LIST OF TABLES

Table 1) NetApp BeeGFS building block configuration profiles 6
Table 2) Hardware requirements. 8
Table 3) Software requirements. 8
Table 4) SuperPOD storage performance guidelines 10
Table 5) SuperPOD storage sizing 10
Table 6) Converged building block capacity sizing 11
Table 7) Storage-only building block capacity sizing 11

Executive summary

Although AI enhances consumers' lives and helps organizations in all industries worldwide to innovate and to grow their businesses, it is a disrupter for IT. To support the business, IT departments are scrambling to deploy high-performance computing (HPC) solutions that can meet the extreme demands of AI workloads. As the race to win with AI intensifies, the need for an easy-to-deploy, easy-to-scale, and easy-to-manage solution becomes increasingly urgent.

The NVIDIA DGX SuperPOD, featuring H100, H200, or B200 GPUs makes supercomputing infrastructure easily accessible for your organization and delivers the extreme computational power that you need to solve even the most complex AI problems. To help you deploy at scale today, this NVIDIA and NetApp turnkey solution removes the complexity and guesswork from infrastructure design and delivers a complete, validated solution including best-in-class compute, networking, storage, and software.

Solution overview

NVIDIA DGX SuperPOD featuring DGX H100, DGX H200, or DGX B200 systems is an AI data center infrastructure platform delivered as a turnkey solution for IT to support the most complex AI workloads facing today's enterprises. It simplifies deployment and management while delivering virtually limitless scalability for performance and capacity. In other words, DGX SuperPOD lets you focus on insights instead of infrastructure. With NetApp EF600 all-flash arrays at the foundation of your NVIDIA DGX SuperPOD, you get an agile AI solution that scales easily and seamlessly. The flexibility and scalability of the solution enable it to support and adapt to evolving workloads, making it a strong foundation to meet your future storage requirements. Servers running BeeGFS are paired with NetApp EF600 storage arrays to form building blocks, giving you a granular approach to growth. By increasing the number of building blocks, you can scale up the performance and capacity of the file system, enabling your solution to manage the most extreme workloads with ease. BeeGFS on NetApp EF600 building blocks connect to DGX H100, DGX H200 and DGX B200 systems through NVIDIA QM9700 InfiniBand switches, ensuring efficient compute and storage operations.

Use case summary

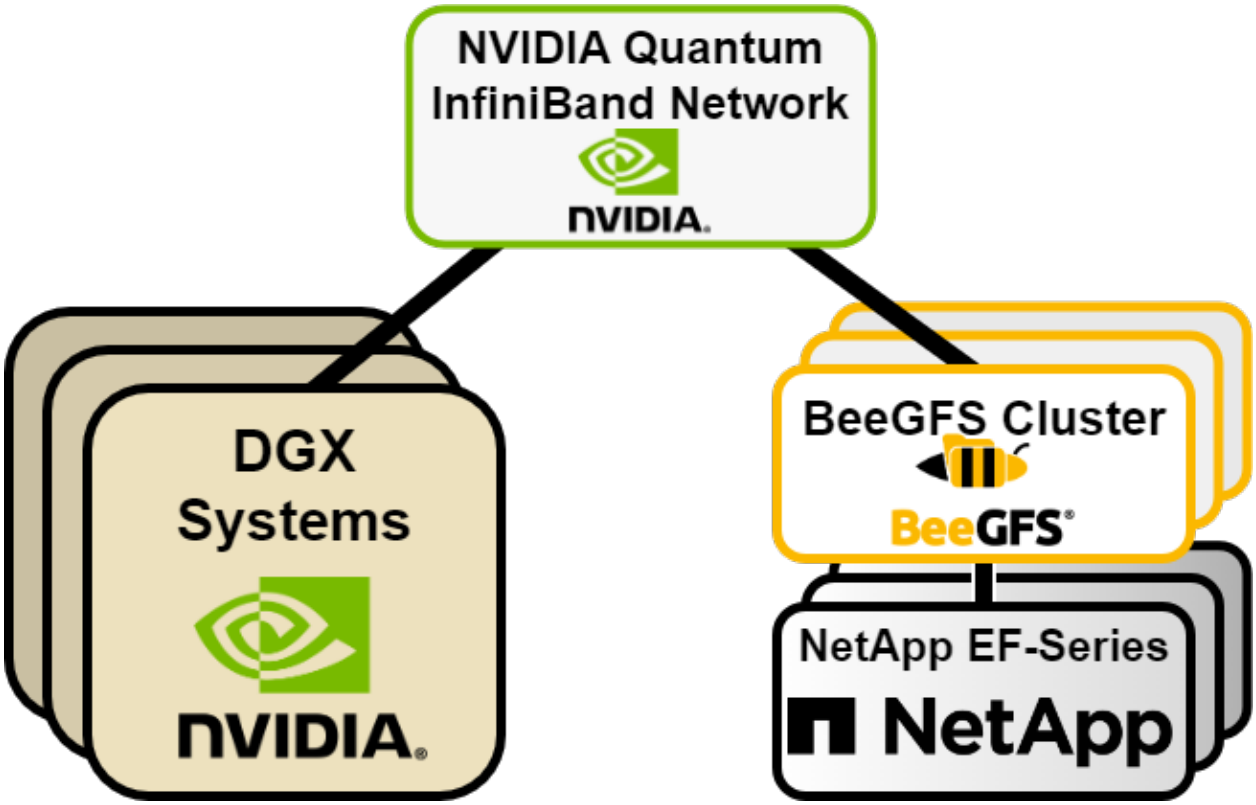
This solution supports the training and inference phases of the AI and DL pipeline. Depending on the application, AI/DL models work with large amounts of different types of data (both structured and unstructured). This difference imposes a varied set of requirements on the underlying storage system, both in terms of size of the data that is being stored and the number of files in the dataset.

The high-level storage requirements include the following:

- The ability to store and to retrieve millions of files concurrently.
- Storage and retrieval of diverse data objects such as images, audio, video, and time-series data.
- Delivery of high parallel performance at low latencies to meet the GPU processing speeds.
- Seamless data management and data services that span the edge, the core, and the cloud.

For the critical training phase of DL, data is typically copied from the data lake into the training cluster at regular intervals. That data is then processed repeatedly by the DL model to achieve the desired ML proficiency. The servers that are used in this phase use GPUs to parallelize computations, creating a tremendous appetite for data. Meeting the raw I/O bandwidth needs is crucial for maintaining high GPU utilizations.

Figure 1) A high-level view of the solution.

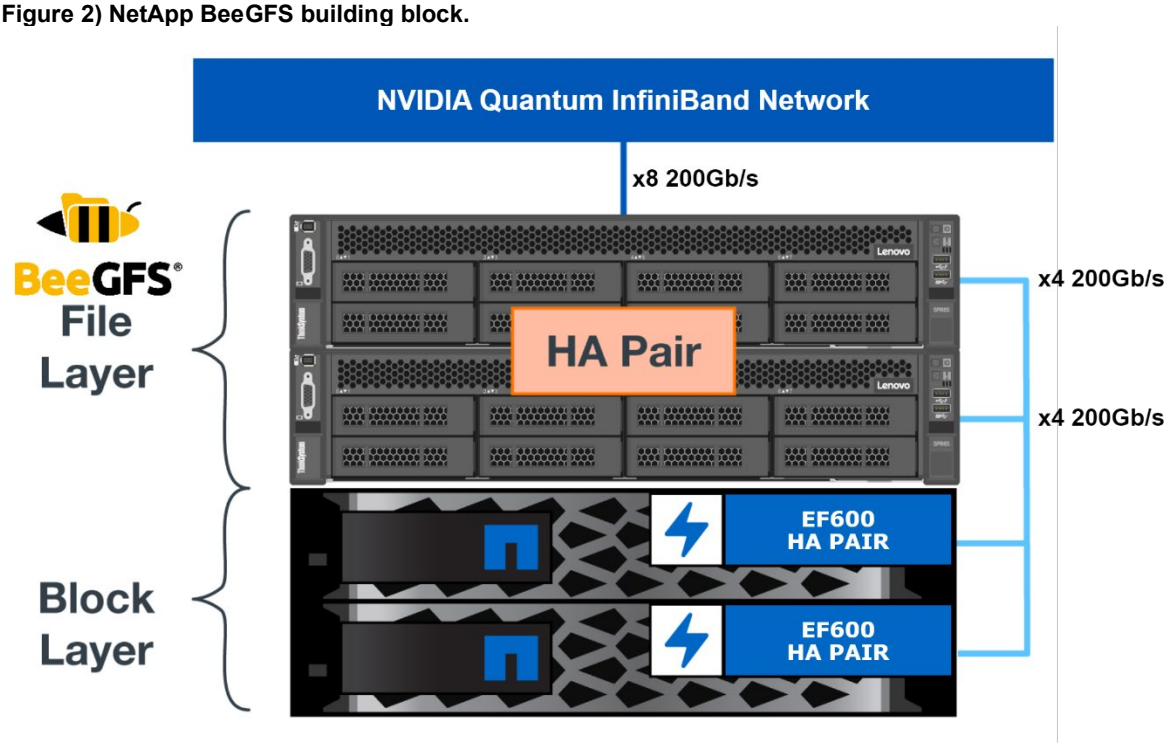


A NetApp BeeGFS building block consists of two NetApp EF600 storage arrays directly attached to two Lenovo ThinkSystem SR665 V3 servers. The BeeGFS file system sits on the Lenovo SR665 V3 servers, known as the file layer, and can be deployed and scaled in different ways depending on storage requirements. For example, use cases primarily featuring numerous small files will benefit from extra metadata performance and capacity, whereas use cases featuring fewer large files might favor more storage capacity and performance for actual file contents. The building block architecture is designed to scale out each of these dimensions. Typically, building blocks are deployed in one of three configuration profiles, as shown in Table 1. The first building block in the BeeGFS cluster uses the base configuration profile, while all other building blocks in the cluster use the converged or storage-only configuration profiles. Figure 2 shows an overview of a single NetApp BeeGFS building block.

Table 1) NetApp BeeGFS building block configuration profiles

Building Block Configuration profile	BeeGFS services
Base	management, metadata, and storage
Converged	metadata and storage
Storage-only	storage

Figure 2) NetApp BeeGFS building block.



High Availability Architecture

BeeGFS on NetApp enhances the functionality of the BeeGFS enterprise edition by integrating seamlessly with NetApp hardware to enable a shared-disk high availability (HA) architecture. In this setup, volumes from the EF600 storage arrays are mapped to both file nodes, allowing them to take over each other's roles as needed. BeeGFS services can be initiated on multiple nodes. Therefore, Pacemaker must ensure that each service and its dependent resources are only running on one node at a time. To maintain this configuration, Pacemaker relies on Corosync to keep the overall cluster state synchronized across all nodes and to establish quorum. With quorum established, Pacemaker can react to failures by restarting BeeGFS resources on another node. However, there are scenarios where Pacemaker might be unable to communicate with the original faulty node to confirm that the resources have stopped. To handle this, Pacemaker uses fencing to isolate the faulty node, ideally by cutting its power, ensuring it is truly down before restarting BeeGFS resources elsewhere. The surviving file node in the HA pair can handle the workload of the faulty node, ensuring near-normal performance even during failures.

Technology requirements

This section covers the technology requirements for the NVIDIA DGX SuperPOD with the BeeGFS on NetApp solution. For a list of technology requirements in DGX SuperPOD, refer to the DGX SuperPOD reference architecture.

NetApp BeeGFS hardware requirements

This solution supports the training and inference phases of the AI and DL pipeline. Depending on the application, AI/DL models work with large amounts of different types of data (both structured and unstructured). This difference imposes a varied set of requirements on the underlying storage system, both in terms of size of the data that is being stored and the number of files in the dataset.

The high-level storage requirements include the following:

- The ability to store and retrieve millions of files concurrently.
- Storage and retrieval of diverse data objects such as images, audio, video, and time-series data.
- Delivery of high parallel performance at low latencies to meet the GPU processing speeds.
- Seamless data management and data services that span the edge, the core, and the cloud.

For the critical training phase of DL, data is typically copied from the data lake into the training cluster at regular intervals. That data is then processed repeatedly by the DL model to achieve the desired ML proficiency. The servers that are used in this phase use GPUs to parallelize computations, creating a tremendous appetite for data. Meeting raw I/O bandwidth needs is crucial for maintaining high GPU utilizations.

Table 2 lists the hardware components required to implement the solution for a single SU. The solution sizing starts with three NetApp BeeGFS building blocks for a single SU. For a complete list of hardware requirements, view the bill of materials.

Table 2) Hardware requirements.

Quantity	Hardware	Components
6	EF600 storage systems	**Memory:** 256GB (128GB per controller)
		Adapter: 2-port 200Gb/HDR (NVMe/IB)
		Drives: 24 Drives, configured to match desired metadata and storage capacity
6	Lenovo SR665 V3 servers	**CPU**: 2x AMD EPYC 9124 16C 3.0GHz
		Memory: 256GB (16x 16GB TruDDR5 4800MHz RDIMM-A)
		PCIe Slots: PCIe Gen5 x16 slots (three per NUMA zone)
1	Ansible Control Node	An Ansible control node is a physical or virtual Linux machine used to manage the cluster.

Software requirements

Table 3 lists the software components used to validate the solution. These software requirements change as new software versions release. For the latest requirements, refer to the NetApp Doc's BeeGFS on NetApp's technical requirements.

Table 3) Software requirements.

Software	Version
SANtricity	11.80.1R3
NVSRAM	N6000-881834-D01.dlp
BeeGFS servers OS	RedHat Enterprise Linux RHEL 9.3
BeeGFS servers Kernel	5.14.0-362.24.1el9_3.x86_64
InfiniBand / RDMA Drivers	MLNX_OFED_LINUX-23.10-3.2.2.0-LT
OpenSM	Opensm-5.17.2 (from MLNX_OFED_LINUX-23.10-3.2.2.0-LTS)
BeeGFS servers multipathing	Device Mapper MultiPath (DMMP)
BeeGFS	7.4.4
Corosync	3.1.5-4
Pacemaker	2.1.4-5

User storage

User storage is provided as an NFS share through the 100 Gb/s Ethernet in-band management fabric. As a key component of NVIDIA's DGX SuperPOD, user storage has multiple functions, including storing home directories, administrative scratch space, log files, and shared storage. Shared storage is essential for NVIDIA's Base Command Manager. NetApp's AFF C-Series platforms offer a consolidated and comprehensive solution for deploying shared user storage over NFS within your DGX SuperPOD environment. The AFF C-Series platforms, such as the C30, connect directly to the 100 Gb/s Ethernet in-band fabric, delivering scalable and reliable user storage for your DGX SuperPOD. For more information, refer to NetApp AFF C-Series – capacity-optimized storage | Netapp.

Data movement

Utilizing NetApp XCP and NetApp Cloud Sync tools on a dedicated data mover node, system administrators can efficiently manage the ingestion and egress of data within a BeeGFS cluster. The data mover node requires access to the in-band network. The performance of XCP transfers scales with the number of CPU cores and the amount of memory available. Therefore, ensure that the data mover node is adequately equipped to handle your data transfer needs. For more information, read Data movement with E-Series and BeeGFS for AI and analytics workflows.

Another option for exporting data from the BeeGFS filesystem is to use NFS. BeeGFS can serve files using NFSv4 by configuring a BeeGFS client host as an NFS server and re-exporting the BeeGFS client mount-point via NFS. Refer to the BeeGFS documentation for configuration instructions.

Power requirements

Power requirements depend on the performance and capacity requirements for a particular installation. A default starting configuration 32 node DGX SuperPOD with the recommended BeeGFS storage blocks requires up to approximately 326,4kW maximum power. See the NVIDIA DGX SuperPOD Data Center Design Reference Guide for a breakdown of thermal and power requirements.

Hardware installation and basic setup

All hardware components should be installed in the data center racks according to the vendor's recommended guidelines. Three building blocks are recommended per DGX SuperPOD SU. Three building blocks can be accommodated in a 42 RU tall rack, with additional room for two DGX systems, and two 1U NVIDIA Quantum-2 InfiniBand switches for the storage network fabric. Specific rack power and cooling capacities determine exactly how many servers can be supported in each rack.

Perform basic setup for each component using the appropriate installation documentation. The following configuration procedures assume that all components have been installed and configured for management access and have been upgraded to the software and/or firmware versions recommended in this validation. For specific details about basic installation and setup, see the appropriate vendor documentation. Links are provided in the deployment procedures section for reference.

Sizing guidelines

The NetApp EF-Series BeeGFS DGX SuperPOD solution includes recommendations for performance and capacity sizing based on verification tests. The objective of a building block architecture is to create a solution that is simple to size by adding building blocks to meet the requirements for a particular BeeGFS system. Using the guidelines below, you can estimate the quantity and types of BeeGFS building blocks needed to meet the requirements of your environment.

Keep in mind that these estimates represent best-case performance. Synthetic benchmarking applications are written and utilized to optimize the use of underlying file systems in ways that real-world applications might not.

Performance sizing

The DGX SuperPOD is designed to support all workloads, but the storage performance required to maximize training performance can vary depending on the type of model and dataset. The guidelines in Table 4 capture three different performance levels required for common AI workloads and dataset sizes, providing guidelines to determine the I/O levels necessary for different types of models.

Table 4) SuperPOD storage performance guidelines

Performance Level	Work Description	Dataset Size	Single SU Aggregate System Read	Single SU Aggregate System Write
Standard	Natural Language Processing, Training Compressed Images, Audio, and Text Data (e.g., LLM Training)	Most datasets can fit within the local compute systems' memory cache during training. The datasets are single modality, and models have millions of parameters.	40 GB/s	20 GB/s
Enhanced	Training with large Video and Image Files, Offline Inference, ETL	Datasets are too large to fit into cache, massive first epoch I/O requirements, workflows that only read the dataset once	125 GB/s	62 GB/s

NetApp recommends sizing the NetApp BeeGFS storage solution to achieve the best storage performance level, starting with three building blocks for a single DGX SuperPOD configuration. The building block architecture scales linearly with each DGX SuperPOD. See Table 5 for sizing guidance for up to four DGX SuperPOD SUs. The building block configuration from the table is a recommendation but can be altered to match storage requirements. Up to five building blocks can be deployed to form a standalone Linux HA cluster. One or more of these standalone BeeGFS HA clusters are combined to create a BeeGFS file system that is accessible to clients as a single storage namespace.

Table 5) SuperPOD storage sizing

DGX SuperPOD Configuration	32 DGX H100	64 DGX H100	96 DGX H100	128 DGX H100
NetApp BeeGFS building blocks	3	6	9	12
HA Cluster Count	1	2	2	3
Building Block Configuration	2 Converged + 1 Storage	3 Converged + 3 Storage	4 Converged + 5 Storage	5 Converged + 7 Storage
Read throughput	197 GB/s	394 GB/s	591 GB/s	789 GB/s
Write throughput	67 GB/s	134 GB/s	202 GB/s	269 GB/s

Capacity sizing for converged building blocks

Table 6 provides recommended capacity sizing for converged (metadata + storage) building blocks. When sizing converged building blocks, you can reduce cost by using smaller drives for metadata volume groups versus storage volume groups.

Table 6) Converged building block capacity sizing

Drive Size (2+2 Raid 1) metadata volume groups	Metadata Capacity (number of files)	Drive size (8+2 RAID 6) storage volume groups	Storage capacity (file content)
1.92TB	1,938,577,200	1.92TB	51.77TB
3.84TB	3,880,388,400	3.84TB	103.55TB
7.68TB	8,125,278,000	7.68TB	216.74TB
15.3TB	17,269,854,000	15.3TB	460.60TB

Capacity sizing for storage-only building blocks

Table 7 provides rule-of-thumb capacity sizing for storage-only building blocks.

Table 7) Storage-only building block capacity sizing

Drive size (10 + 2 RAID 6) storage volume groups	Storage capacity (file content)
1.92TB	59.89TB
3.84TB	119.80TB
7.68TB	251.89TB
15.3TB	538.55TB

DGX SuperPOD network fabrics

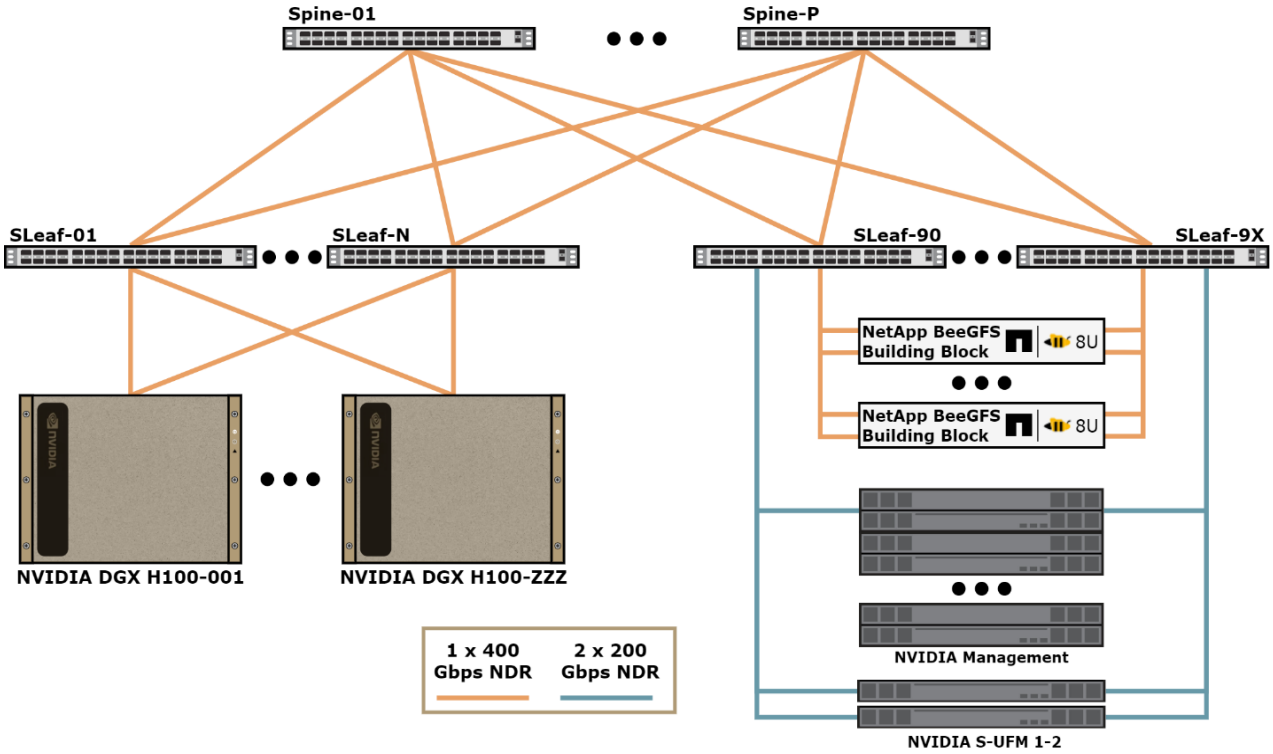
DGX SuperPOD reference architecture includes four network fabrics. Here is a high-level overview of each network fabric:

- **Compute fabric:** A highly optimized fabric connecting all compute nodes within a DGX SuperPOD.
- **Storage fabric:** An NVIDIA Quantum-2 InfiniBand network fabric to provide connectivity between compute nodes, NetApp BeeGFS building blocks, and management nodes.
- **In-band management network:** A high-speed Ethernet fabric that connects compute nodes, management nodes, and storage nodes, enabling cluster management services, access to the data NFS tier, connectivity to in-cluster services, and external services.
- **Out-of-band management network:** A secure and isolated Ethernet fabric for management access to all devices in the DGX SuperPOD configuration.

The NetApp BeeGFS storage solution connects to the storage, in-band, and out-of-band network fabrics. For more information, refer to the DGX SuperPOD reference architecture.

The storage fabric is an NVIDIA Quantum-2 InfiniBand network, designed to deliver high-bandwidth connectivity among compute nodes, NetApp BeeGFS building blocks, and management nodes. This fabric utilizes NVIDIA Quantum-2 QM9700 InfiniBand switches, which feature 64 NDR 400 Gb/s ports derived from 32 OSFP ports within a 1U form factor. Figure 3 provides an overview of the DGX SuperPOD storage network architecture.

Figure 3) DGX SuperPOD storage network architecture



NetApp BeeGFS building block connections

Figures 4 and 5 provide an overview of the connectivity for three NetApp BeeGFS building blocks cabled to the DGX SuperPOD storage fabric. Each Lenovo SR665 V3 server is equipped with four NVIDIA ConnectX-7 InfiniBand host channel adapters (HCA) for storage connectivity. To maximize PCIe bidirectional bandwidth, the **iXa** port of each HCA connects to the NetApp storage arrays, while each **iXb** port connects to the storage network. Each building block connects to the QM9700 storage network switch using four NDR (400Gb/s) to HDR (200Gb/s) InfiniBand splitter cables. Two of these splitter cables branch out and connect to the ports highlighted in light green, and the other two cables connect to the ports highlighted in dark green. This cabling scheme is designed to ensure high availability, optimize server CPU performance, and maximize PCIe bandwidth.

Note: It is not necessary to cable the solution to two separate storage switches, but it provides are extra layer of redundancy.

Figure 4) Three building blocks cabled to the DGX SuperPOD storage fabric

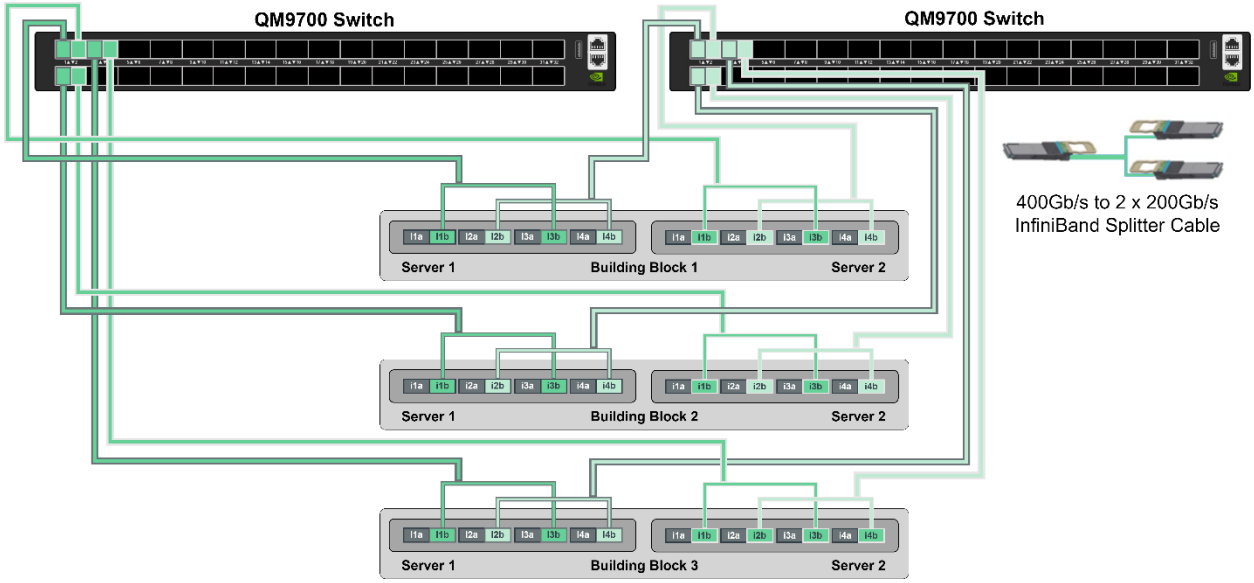
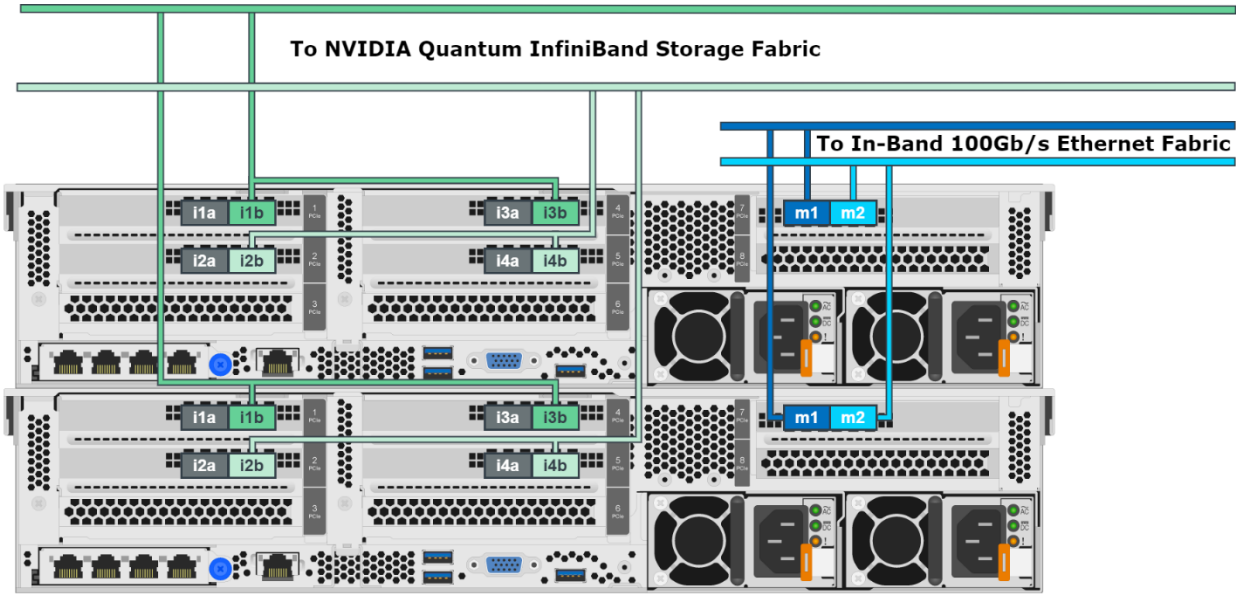
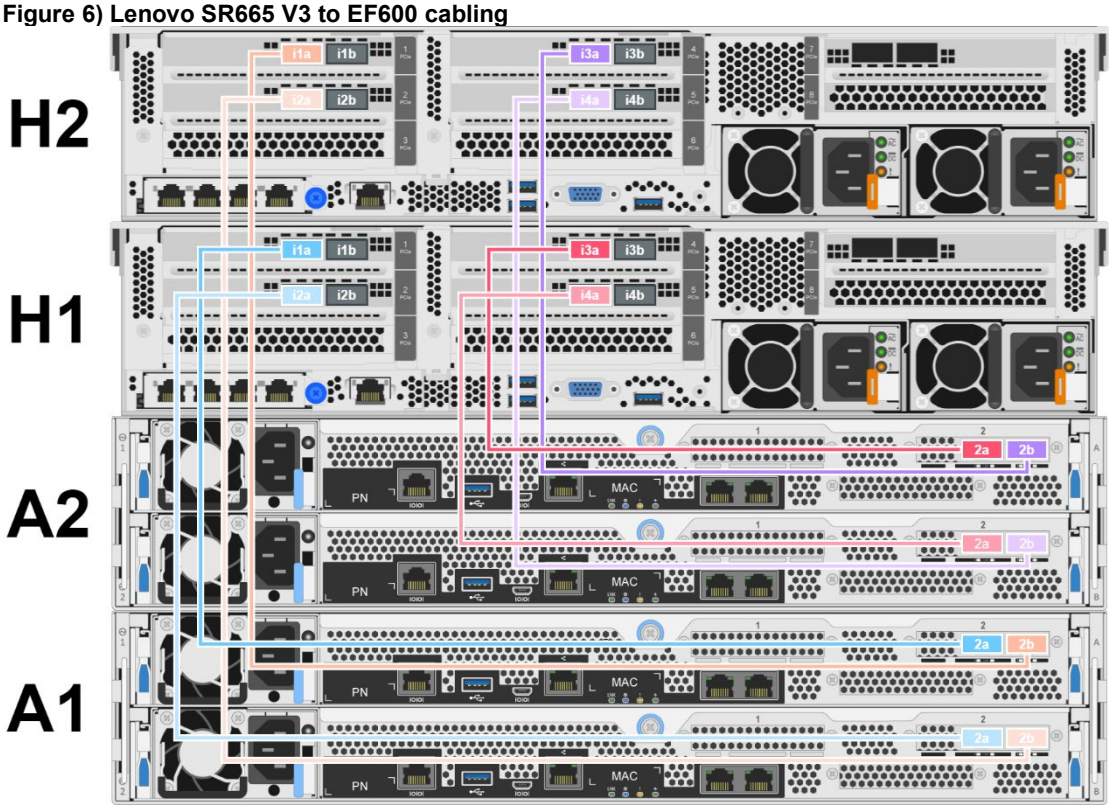


Figure 5) Lenovo SR665 V3 to network cabling



For each building block, the Lenovo SR665 V3 servers are directly connected to two EF600 storage arrays using eight HDR InfiniBand cables. Figure 6 shows the cabling topology for a building block.

Figure 6) Lenovo SR665 V3 to EF600 cabling



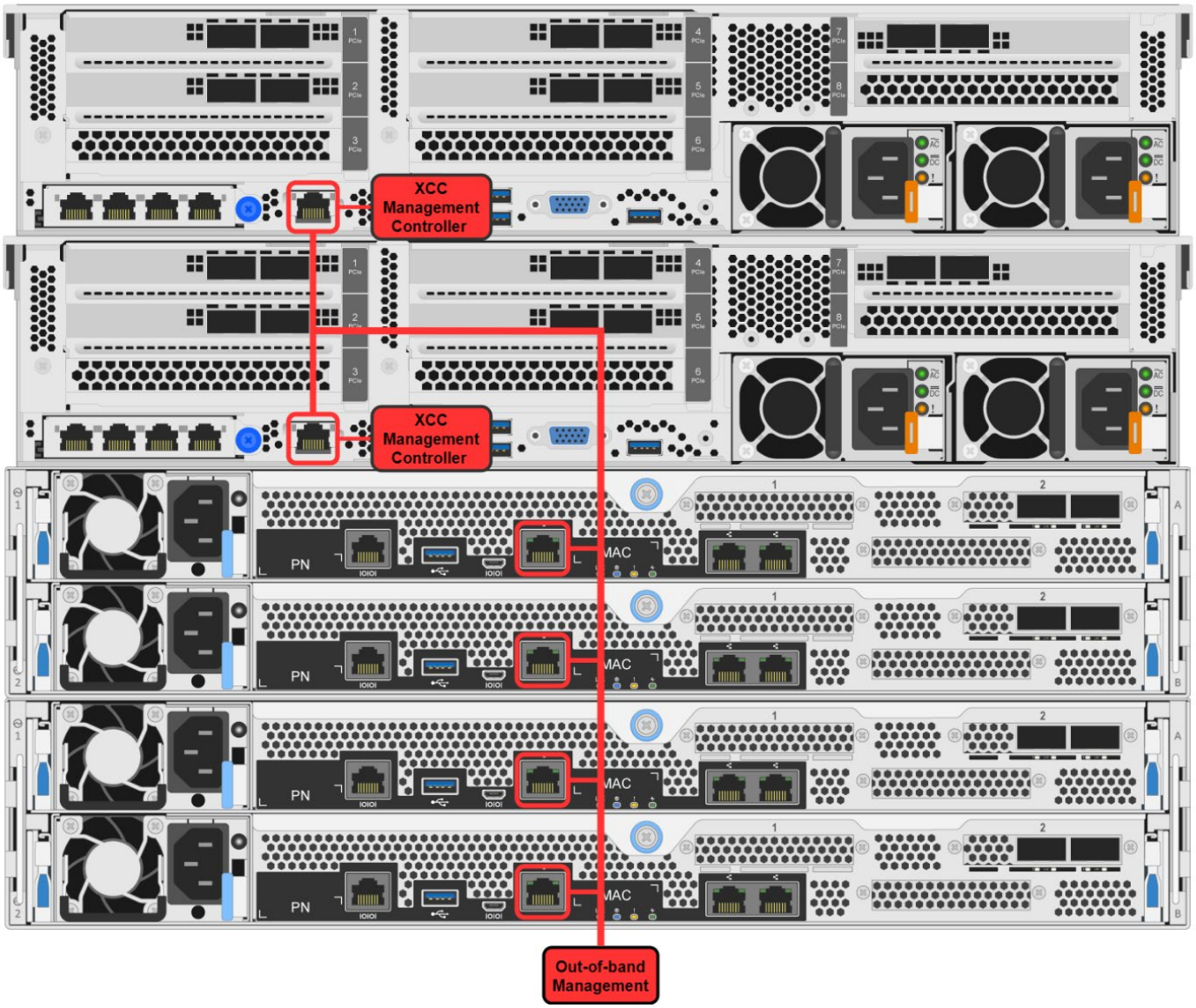
In-band management network

The in-band network is an Ethernet fabric integrated to provide a communication layer for several key functions within DGX SuperPOD components. Figure 5 shows how each NetApp BeeGFS building block, equipped with two ConnectX-6 Ethernet NICs, connects to the in-band network fabric using four 100GbE cables. Building blocks leverage the network for functions such as data movement into the BeeGFS file system, storage node management, and access to the data NFS tier.

Out-of-band management network

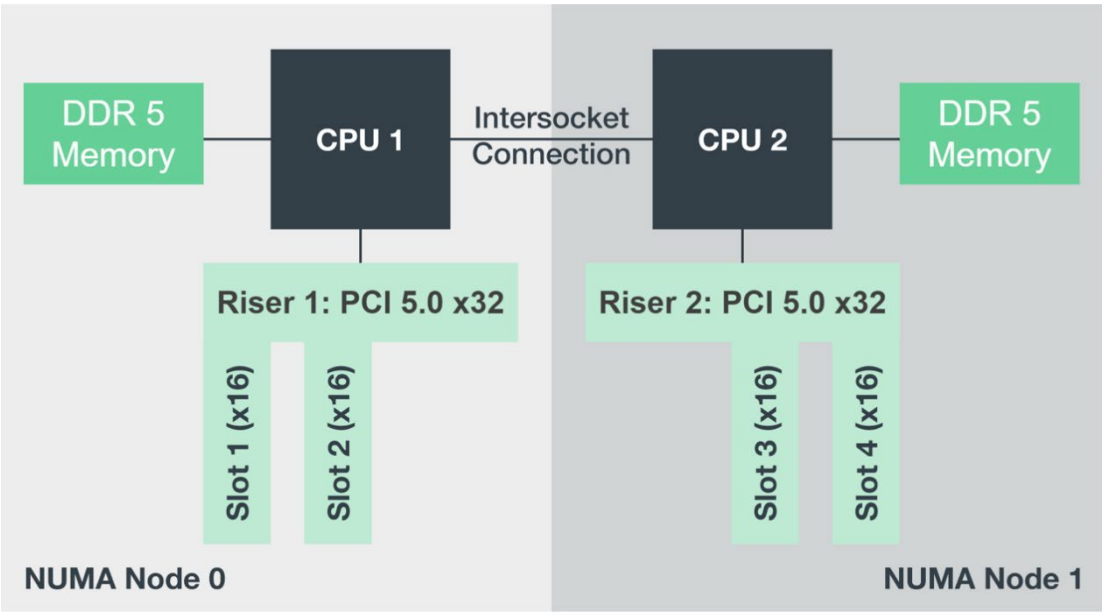
The out-of-band (OOB) management network is a secure and isolated Ethernet fabric for management access to all devices in the DGX SuperPOD configuration. Each node in a building block requires OOB connections for management, Ansible deployment, and high-availability service communications. Refer to Figure 7 for OOB management port connectivity. The XCC management controller ports integrate with high-availability services and provide an option as a fencing agent.

Figure 7) Building block OOB management



Lenovo SR665 V3 servers have two CPU sockets configured as separate NUMA zones, which include local access to an equal number of PCIe slots and memory. NVIDIA ConnectX-7 InfiniBand adapters must be populated in the appropriate PCI risers or slots, so the workload is balanced over the available PCIe lanes and memory channels. You balance the workload by fully isolating work for individual BeeGFS services to a particular NUMA node. The goal is to achieve similar performance from each file node as if it were two independent single socket servers.

Figure 8) File node NUMA configurations

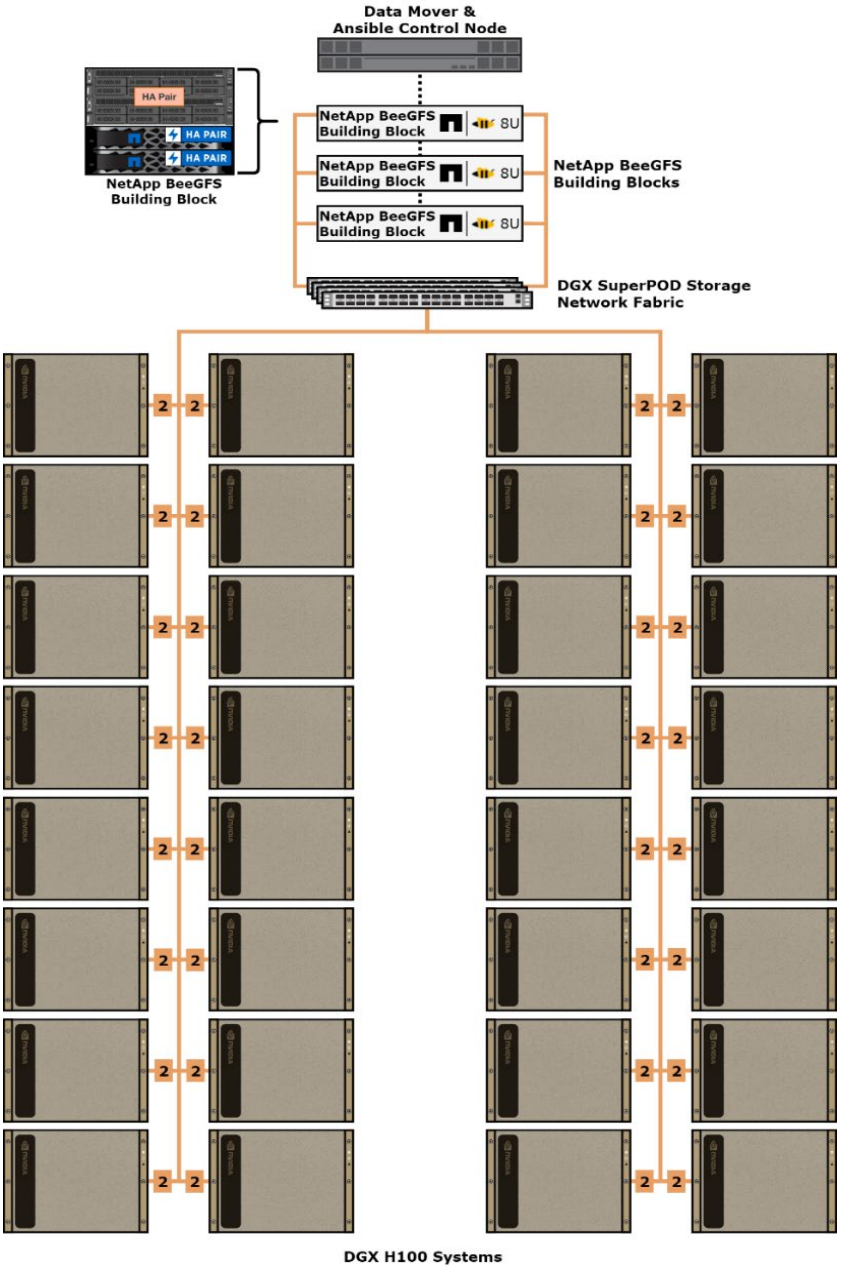


Designs and Architectures

Single DGX SuperPOD SU with three building blocks

The architecture shown in Figure 9 details the storage InfiniBand fabric for a single DGX SuperPOD scalable unit (SU) comprised of 32 DGX H100 systems with three EF600 building blocks. Each DGX H100 system connects to the InfiniBand storage fabric with two NDR 400Gb/s links, for a total of 64 links in each SU. Each EF600 building block connects to the InfiniBand storage fabric using 4 NDR (400Gb/s) to HDR (200Gb/s) splitter cables, resulting in a total of 12 NDR links for three building blocks. The 1GbE out-of-band management network is utilized by the Ansible control node and high availability services to perform management and communication tasks.

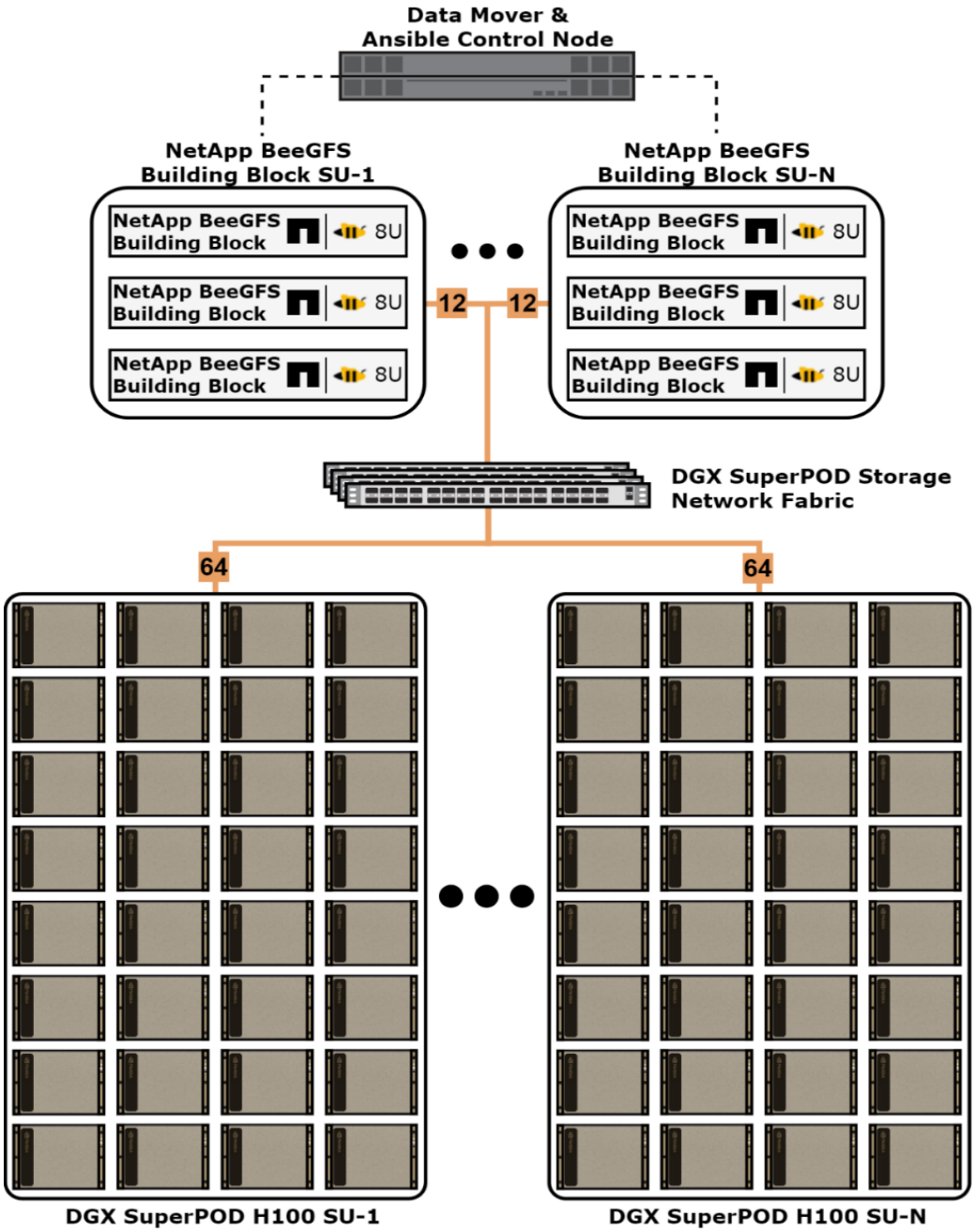
Figure 9) Single DGX SuperPOD with three building blocks storage architecture



Scaling DGX SuperPOD with building blocks

The architecture shown in Figure 10 details the storage InfiniBand fabric for any number of DGX SuperPOD scalable units (SU) with building blocks. Each DGX H100 system connects to the InfiniBand storage fabric with two NDR 400Gb/s links, for a total of 64 links per SU. Each EF600 building block connects to the InfiniBand storage fabric using 4 NDR (400Gb/s) to HDR (200Gb/s) splitter cables, resulting in a total of 12 NDR links for three building blocks. The architecture scales linearly with each DGX SuperPOD SU requiring three building blocks. The out-of-band management network is utilized by the Ansible control node and high availability services to perform management and communication tasks.

Figure 10) Scaling DGX SuperPOD with EF600 building blocks storage architecture



Deployment procedures

This section outlines the steps for deploying and configuring the NetApp BeeGFS storage solution within a DGX SuperPOD environment. While this guide includes additional procedures tailored for the DGX SuperPOD setup, please note that it complements the comprehensive BeeGFS with NetApp E-Series documentation available online.

Deploying NetApp storage systems for use with DGX H100, DGX H200, or DGX B200 systems involves the following tasks:

1. **Ansible control node setup**
2. **Hardware deployment**
3. **BeeGFS deployment and configuration**
4. **DGX H100, DGX H200, or DGX B200 client configuration**

DGX SuperPOD deployment and configuration using Ansible

The DGX SuperPOD on NetApp with BeeGFS solution is recommended to be deployed using Ansible, which is a popular IT automation engine used to automate application deployments. Ansible uses a series of files collectively known as an inventory, which models the BeeGFS file system you want to deploy.

Ansible allows companies such as NetApp to expand on built-in functionality using collections on Ansible Galaxy (see NetApp E-Series BeeGFS collection). This automated approach reduces the time needed to deploy the BeeGFS file system and the underlying HA cluster. In addition, it simplifies adding building blocks to expand the existing file systems.

Because numerous steps are involved in deploying BeeGFS on NetApp solution, NetApp does not support manually deploying the solution.

Ansible control node setup

To set up an Ansible control node, you must identify a virtual or physical machine with network access to the management ports of all EF600 storage arrays and BeeGFS nodes. The control node will be used to configure the solution.

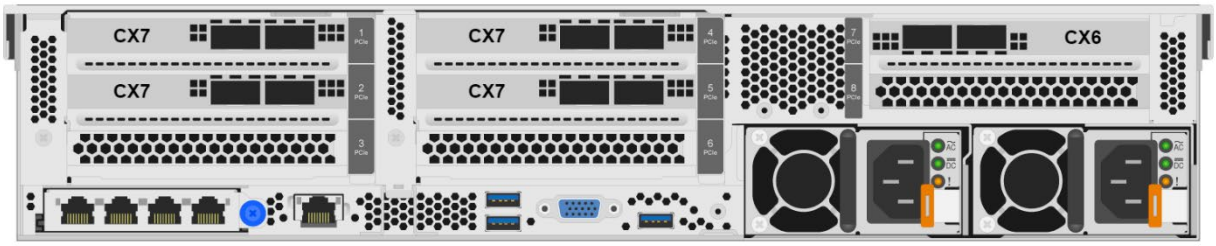
See Set up an Ansible control node for setup instructions.

Hardware deployment

The following steps are identical for each building block in the cluster, regardless of whether it is used to run both BeeGFS metadata and storage services, or just storage services.

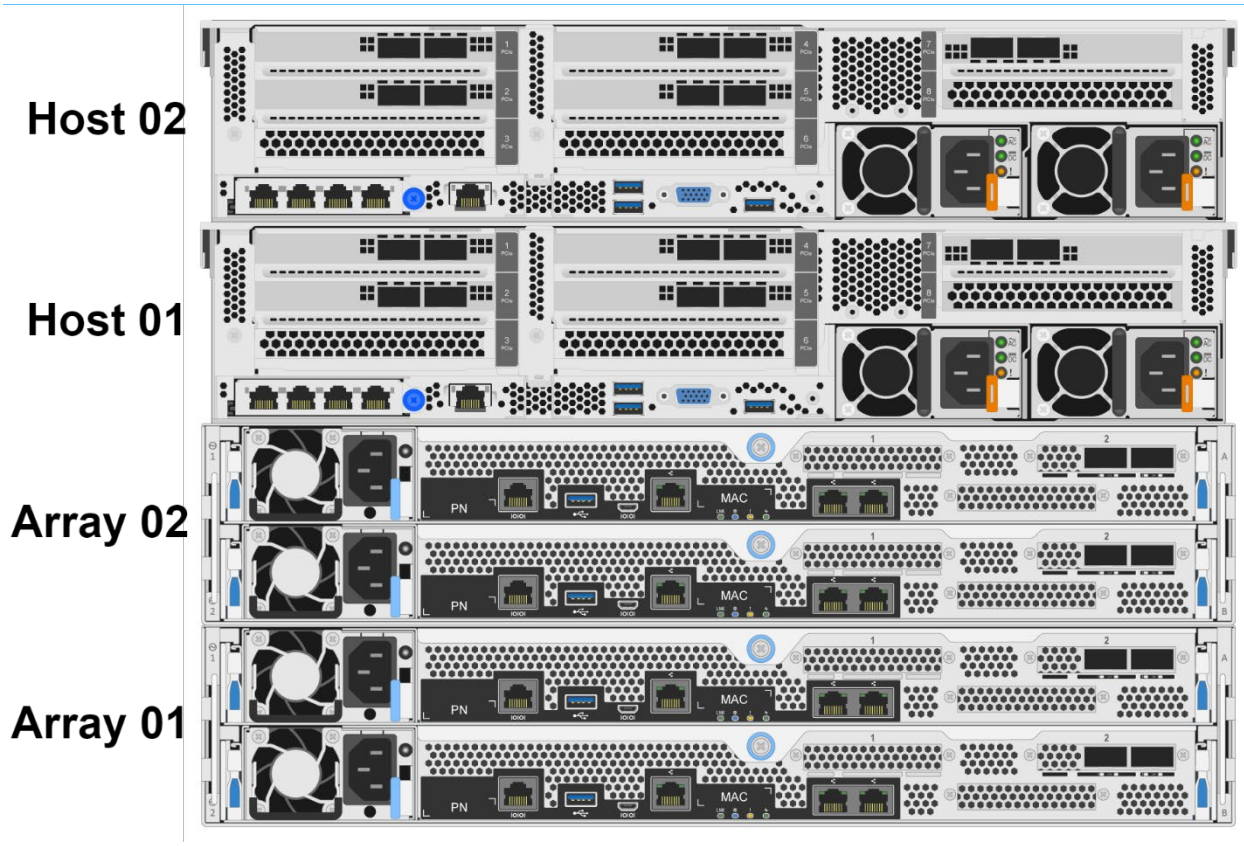
1. Set up each Lenovo SR665 V3 with the appropriate HCAs. Figure 11 can be used as reference:
    a. Insert four MCX755106AS-HEAT adapters into PCIe slots 1, 2, 4, and 5.
    b. Insert one MCX653106A-ECAT adapter into PCIe slot 7.

Figure 11) Lenovo SR665 V3 host card adapter layout



2. Rack the building blocks so the two Lenovo SR665 V3 servers are above the EF600 storage arrays. Figure 12 shows the correct hardware configuration for the BeeGFS building block using Lenovo ThinkSystem SR665 V3 servers as the file nodes (rear view).
   For detailed procedures, see Deploy hardware.

Figure 12) Lenovo SR665 V3 host card adapter layout



3. If not already done, install the 24 drives in each of the EF600 storage arrays.
4. For each building block, connect the EF600 storage arrays to the Lenovo server's ConnectX-7 cards using the 1m InfiniBand HDR 200Gb direct attach copper cables, so that they match the topology shown in figure 6 of the NetApp BeeGFS building block connections section.
5. For each building block, connect remaining ports of the Lenovo server's ConnectX-7 cards to the DGX SuperPOD storage network using the 400Gb/s to 2x 200Gb/s InfiniBand splitter cables, so that they match the topology shown in figure 4 of the NetApp BeeGFS building block connections section.
   For each building block, connect the Lenovo servers to the DGX SuperPOD in-band network using the ConnextX-6 ethernet adapters with 100GbE cables. Refer to figure 5 of the NetApp BeeGFS building block connections section for assistance.

Software deployment

The deployment procedures cover the following configuration profiles:

- One base building block that includes management, metadata, and storage services.
- A second building block that includes metadata and storage services.
- A third building block that includes only storage services.

These profiles demonstrate the full range of recommended configuration profiles for the NetApp BeeGFS building blocks. For each deployment, the number of metadata and storage building blocks or storage services-only building blocks may vary in the procedures, depending on capacity and performance requirements. Refer to the following links for instructions on deploying the BeeGFS cluster.

1. **Set up file and block nodes.**
2. **Tune system settings for performance.**
3. **Create the Ansible inventory.**
4. **Define Ansible inventory for BeeGFS building blocks.**
5. **Deploy BeeGFS using Ansible.**

BeeGFS client installation and configuration

Mounting BeeGFS to the DGX H100, DGX H200, or DGX B200 systems requires installing the BeeGFS client package by using the apt package manager. The BeeGFS client automatically handles the compiling of the BeeGFS kernel module that provides a normal mountpoint, which ensures applications can access BeeGFS without modification. This module can be installed on all supported Linux kernels without the need for patches.

Before you start the BeeGFS client, make the following minor changes:

1. Update the `buildArgs` line in `/etc/beegfs/beegfs-client-autobuild.conf` to ensure the client module is built with support for remote direct memory access (RDMA). If the OpenFabrics Enterprise Distribution (OFED) kernel modules are in use, you must supply the path to the OFED kernel modules. For the OFED version used on the DGX A100 systems, the following configuration is a result:

```
buildArgs=-j8 BEEGFS_OPENTK_IBVERBS=1 OFED_INCLUDE_PATH /usr/src/ofa_kernel/5.3.0-59-
generic/include/
```

2. In `/etc/beegfs/beegfs-client.conf`, set `sysMgmtdHost` to the IP address of the server running the BeeGFS management service. The following parameters are overridden:
   - `connMaxInternodeNum = 12 > 128`
   - `connRDMABufNum = 70 > 32`
   - `connRDMABufSize = 8192 > 65536`
3. Start or enable BeeGFS `beegfs-helperd` and `beegfs-client services` by using `systemctl`:

```
systemctl start beegfs-helperd beegfs-client
systemctl enable beegfs-helperd beegfs-client
```

The BeeGFS file system is now accessible at the default `/mnt/beegfs` mount point.

4. Using Base Command Manager (BCM) a golden image of DGX compute node configured with the BeeGFS mount point can be distributed across the remaining DGX compute nodes.

Optimizing performance with BeeGFS

One of the benefits of BeeGFS is the ability to optimize for multiple types of workloads that might need to share the same global file system. Specifically, BeeGFS allows configuration of the “stripe pattern” used to split each file to be written/read in parallel across all of the BeeGFS storage services and targets (in this instance, NetApp volumes). According to the BeeGFS Documentation website:

Striping in BeeGFS can be configured on a per-directory and per-file basis. Each directory has a specific stripe pattern configuration, which will be derived to new subdirectories and applied to any file created inside a directory. There are currently two basic parameters that can be configured for stripe patterns:

- desired number of storage targets for each file
- chunk size (or block size) for each file stripe

Striping can be easily applied and updated post deployment, therefore, different patterns were used to optimize each benchmark test. For more information about the stripe pattern used for each test ID, see NVA-1156-DESIGN: NetApp EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS Design Guide. To configure striping in BeeGFS, run the following command:

```
beegfs-ctl -setpattern -numtargets=16 -chunksize=2m /mnt/beegfs
```

Solution verification

This solution was validated by using both synthetic storage benchmarks and MLPerf v0.7 training benchmark. For more information about the validation test results, see NVA-1156-DESIGN: NetApp EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS Design Guide.

FIO storage throughput benchmark test

To perform the FIO throughput benchmark test, run the following commands to create the test files and perform I/O to them:

```
/usr/bin/fio --create_only=1 --rw=write --direct=1 --ioengine=posixaio -- iodepth=32 --
create_serialize=0 --fallocate=none --group_reporting=1 -- disable_lat=1 --disable_clat=1 --
disable_slat=1 --startdelay=5 --ramp_time=3 --runtime=180 --time_based=1 --invalidate=1 --
blocksize=1024k -- size=4194304k --numjobs=120 --
directory=/mnt/fs_under_test/fiodir.20200629205829 /usr/bin/fio --rw=write --direct=1 --
ioengine=posixaio --iodepth=32 -- create_serialize=0 --fallocate=none --group_reporting=1 --
disable_lat=1 -- disable_clat=1 --disable_slat=1 --startdelay=5 --ramp_time=3 --runtime=180 - -
time_based=1 --invalidate=1 --blocksize=1024k --size=4194304k --numjobs=120 --
directory=/mnt/fs_under_test/fiodir.20200629205829 /usr/bin/fio --rw=read --direct=1 --
ioengine=posixaio --iodepth=32 -- create_serialize=0 --fallocate=none --group_reporting=1 --
disable_lat=1 -- disable_clat=1 --disable_slat=1 --startdelay=5 --ramp_time=3 --runtime=180 - -
time_based=1 --invalidate=1 --blocksize=1024k --size=4194304k --numjobs=120 --
directory=/mnt/fs_under_test/fiodir.20200629205829
```

FIO storage IOPS benchmark test

To perform the FIO IOPS benchmark test, run the following commands:

```
/usr/bin/fio --create_only=1 --rw=write --direct=1 --ioengine=posixaio -- iodepth=32 --create_serialize=0 --
fallocate=none --group_reporting=1 -- disable_lat=1 --disable_clat=1 --disable_slat=1 --startdelay=5 --
ramp_time=3 --runtime=180 --time_based=1 --invalidate=1 --blocksize=1024k -- size=4194304k --numjobs=180 --
directory=/mnt/fs_under_test/fiodir.20200629213502 /usr/bin/fio --rw=write --direct=1 --ioengine=posixaio --
iodepth=32 -- create_serialize=0 --fallocate=none --group_reporting=1 --disable_lat=1 -- disable_clat=1 --
disable_slat=1 --startdelay=5 --ramp_time=3 --runtime=180 - -time_based=1 --invalidate=1 --blocksize=4k --
size=4194304k --numjobs=180 -- directory=/mnt/fs_under_test/fiodir.20200629213502 /usr/bin/fio --rw=read --
direct=1 --ioengine=posixaio --iodepth=32 -- create_serialize=0 --fallocate=none --group_reporting=1 --
disable_lat=1 -- disable_clat=1 --disable_slat=1 --startdelay=5 --ramp_time=3 --runtime=180 - -time_based=1 --
invalidate=1 --blocksize=4k --size=4194304k --numjobs=180 -- directory=/mnt/fs_under_test/fiodir.20200629213502
```

MLPerf Training v0.7 benchmark test

This test was performed according to the configuration defined in the MLPerf Training v0.7 benchmark specifications. The configuration of this benchmark is outside the scope of this document. For more information about using MLPerf to perform this and other DL benchmark tests, see the MLPerf training overview.

Conclusion

For AI, ML/DL training, and HPC infrastructure, the NVIDIA DGX SuperPOD, featuring DGX H100, DGX H200, or DGX B200 systems, NetApp EF600 storage systems, the BeeGFS parallel file system provides next-generation storage and data management platforms required for these types of workloads. By using BeeGFS with EF600 storage systems, this verified architecture can be implemented at almost any scale, from a single DGX system paired with a single BeeGFS building block, up to potentially 128 DGX H100, DGX H200, or DGX B200 systems with a scalable number of BeeGFS building blocks, all presenting a single storage namespace.

Bill of materials (BOM) for 1 SU

Part Number	Product Description	Quantity
NetApp E-Series EF600 storage systems		
EF600A-128GB-C	EF600 128GB Memory Storage Controller	12
EF600-NVME-IB-PER	EF600 Personality, NVMe-IB HIC	6
X-56038-00-C	2-port 200Gb/HDR NVMe/IB HIC	12
E-X5725A-C	24 Drive NVMe SSD Enclosure	6
E-X4141B-C, or E-X4140B-C, or E-X4137B-C	Storage System NVMe SSDs **E-X4141B-C**: SSD,15.3TB,NVMe,SED,NE224,-C, or **E-X4140B-C**: SSD,7.6TB,NVMe,SED,NE224,-C, or **E-X4137B-C**: SSD,3.8TB,NVMe,SED,NE224,-C	144
EF600-OS		
OS-SANTRICITY-NVME-01	OS Enable, Per-0.1TB, SANTRCTY, Low-Latency,01	7344
Services		
CS-BASE-SUPPORT	Base Software Support	6
CS-A2-4R	SupportEdge Premium 4hr Onsite	6
PS-DEPLOY-STAND-EF-SERIES	PS Deployment, Standard, EF-Series	6

Part Number	Product Description	Quantity
Additional Parts-Hardware		
X6561-R6	Cable, Ethernet, 2m RJ45 CAT6	12
Lenovo ThinkSystem SR665 V3 Server (or other verified server)	**Processors:** 2x AMD EPYC 9124 16C 3.0 GHz (configured as two NUMA zones). **Memory:** 256GB (16x 16GB TruDDR5 4800MHz RDIMM-A) **Riser Cards**: 2x BPQV, 1x BLL9 **PCIe Expansion:** Six PCIe Gen5 x16 slots (three per NUMA zone) Product Guide	6
MCX755106AS-HEAT	NVIDIA ConnectX-7 HCA, NDR-200/200Gb IB/EN, QSFP112, 2-port, PCIe Gen5 x16, InfiniBand Adapter. 4 per server for InfiniBand storage network.	24
MCX653106A-ECAT	NVIDIA ConnectX-6 NIC, 100GbE, QSFP56, 2-port, PCIe Gen3/4 x16, VPI Adapter. 1 per server for in-band Ethernet management network	6
Ansible control node	A physical or virtual machine with network access to NetApp storage arrays and BeeGFS nodes.	1
Additional Parts-OS		
SW-SSP-BEEGFS-UNLTSV-LT	NetApp BeeGFS Support License (Unlimited-Target Server)	6
	RHEL Server Physical, 2 Skt Standard Subscription	6
	RHEL High Availability, 2 Skt Subscription	6
	BeeGFS Enterprise Edition License	6
Cables		
MCP1650-H001E30	NVIDIA passive copper InfiniBand cable, QSFP56, 200Gb/s. 4 per server direct connection from block nodes to file nodes.	24
MFA1A00-C0XX	NVIDIA active Ethernet cable, QSFP, 100GbE. 2 per server for in-band Ethernet management network. Length will vary.	12
Storage network connections provided by NVIDIA	NVIDIA single port transceiver, QSFP112, 400Gb/s, MPO12 APC. 2 per server.	12
	NVIDIA passive fiber cable, MMF, MPO12 APC to 2xMPO12 APC. 2 per server.	12

Where to find additional information

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

NetApp EF-Series systems

- NVA-1156-DESIGN: NetApp EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS Design Guide
  https://www.netapp.com/pdf.html?item=/media/25445-nva-1156-design.pdf
- NetApp EF-Series product page
  https://www.netapp.com/data-storage/ef-series/
- EF600 datasheet
  https://www.netapp.com/pdf.html?item=/media/19339-DS-4082.pdf
- NetApp AI and HPC solutions
  https://www.netapp.com/artificial-intelligence/high-performance-computing/
- BeeGFS with NetApp E-Series Solution Deployment
  https://docs.netapp.com/us-en/beegfs/index.html

NetApp Interoperability Matrix Tool

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

NVIDIA DGX systems

- NVIDIA DGX SuperPOD Architecture
  https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/index.html#
- NVIDIA DGX B200 Systems
  https://www.nvidia.com/en-us/data-center/dgx-b200
- NVIDIA DGX H200/H100 Systems User Guide
  https://docs.nvidia.com/dgx/dgxh100-user-guide/
- NVIDIA GPU Cloud
  https://www.nvidia.com/en-us/gpu-cloud/

NVIDIA Networking

- NVIDIA Quantum 2 QM9700 IB switch
  https://docs.nvidia.com/networking/display/qm97x0pub/introduction

Machine learning frameworks

- TensorFlow: An Open-Source Machine Learning Framework for Everyone
  https://www.tensorflow.org/
- Horovod: Uber's Open-Source Distributed Deep Learning Framework for TensorFlow
  https://eng.uber.com/horovod/
- Enabling GPUs in the Container Runtime Ecosystem
  https://devblogs.nvidia.com/gpu-containers-runtime/

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 IMT lists the supported combinations of NetApp product components and versions, enabling you to build a NetApp-supported configuration. Keep in mind that the actual performance and compatibility outcomes will vary based on your specific installation and adherence to NetApp's published specifications.

NVA-1174-DEPLOY