Technical Report

Entry-level HPC systems with NetApp E-Series and IBM Spectrum Scale

Reference architecture

Jochen Zeller, SVA; Jürgen Türk, NetApp; and Ulf Troppens, IBM
May 2021 | TR-4884

In partnership with

Abstract

This paper presents a reference architecture for entry-level HPC systems based on NetApp® E-Series storage systems and IBM Spectrum® Scale.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Solution overview

These days, many organizations are required to process data acquired from cameras, genome sequencers, super microscopes, and other devices. Laptops and single workstations are too small to process those datasets. However, high-performance computing (HPC) technology has proven to be a cost-effective foundation to acquire, manage, and store large datasets, and to analyze them quickly. Faster access to analysis results in a competitive advantage.

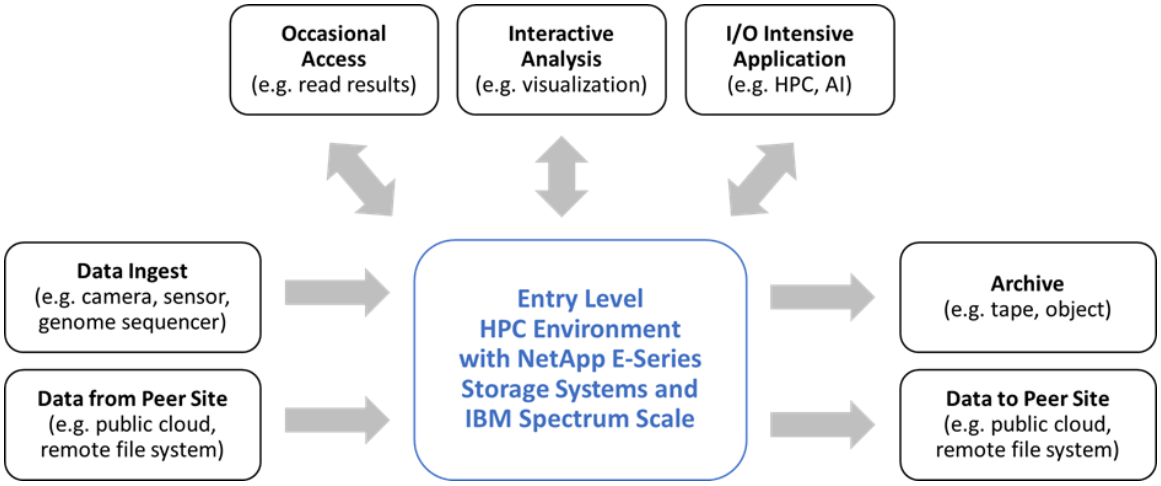
This paper presents a reference architecture for entry-level HPC systems based on NetApp E-Series storage systems and IBM Spectrum Scale. Use cases for this reference architecture include:

- Entry-level HPC and HPC-like workloads, including artificial intelligence (AI) and analytics
- Consolidation of workstations
- Replacement of NFS or SMB file service for data-intensive science and engineering

Solutions for this market segment are typically extremely cost constrained: Customers have a limited budget for storage and a capacity requirement. This reference architecture addresses both dimensions by integrating NL-SAS drives to meet the capacity requirement and SSD drives to boost performance.

Deployments of the reference architecture are typically embedded in a workflow, where data is acquired from instruments or project partners (Figure 1). The locally stored data is then accessed, explored, and analyzed. Finally, data and analysis results are archived or shared with project partners.

Figure 1) System context.



Solution components

In this section we introduce the off-the-shelf components that we used. In the next section we describe how to integrate those components to build the reference architecture for entry-level HPC systems.

Composable infrastructure

Users of HPC systems have varying performance and functional needs. For instance, some users need to run simulations where all data is created by the simulation itself, while other users need to analyze data acquired from different kinds of instruments. Both generated and acquired data quickly add up to hundreds of terabytes or a few petabytes of data that need to be kept online for iterative processing.

There are also differences in the application workload. For cutting-edge science and engineering and new kinds of datasets, users typically start with the interactive exploration of data to plan the subsequent I/O intensive analysis of that data. Analysis can include traditional HPC applications as well as new data analysis techniques such as Hadoop and Spark, machine learning (ML), and deep learning (DL).

Most customers have an existing infrastructure such as servers and networks, authentication services, monitoring, and tooling for deploying and configuring operating systems and applications. Entry-level HPC systems must be integrated into customers' existing infrastructure and services to protect their investments.

A flexible solution architecture is required to support such varying and evolving requirements. The reference architecture for entry-level HPC systems provides composable building blocks that can be customized for varying workload and resilience requirements:

- **Compute.** Modern HPC systems are built on off-the-shelf, rack-mounted servers. They typically have one or two CPUs per server, mostly x86_64 (Intel, AMD) or POWER® (IBM, OpenPOWER), and are optionally configured with GPUs or FPGAs to accelerate selected workloads, sometimes pre-configured such as Nvidia DGX systems. The servers are typically equipped with 256GB to 1TB of memory (RAM).
- **Storage.** Storage systems provide the storage capacity for generated and acquired data, analysis results, application code, and other datasets. For this reference architecture, NetApp E-Series systems was chosen because they allow flexible integration of SSDs and NL-SAS drives in the same storage system and they meet the performance requirements of entry-level HPC workloads.
- **Software-defined infrastructure.** HPC systems include a software layer to simplify and optimize the use and management of the available compute and storage resources. For this reference architecture, IBM Spectrum Scale was chosen as the file system because it makes the underlying storage capacity of the NetApp E-Series systems available as highly scalable, high-performance file and object storage. The reference architecture supports configuring an optional job scheduler to simplify and optimize the use and management of the compute resources.
- **Dedicated data network.** Based on experience in the field, it is an industry best practice to provide a dedicated high-speed network that connects all components of an HPC system. In addition, this network should be configured and not be connected to existing shared data networks such as a data center network or a campus network. This reference architecture supports building the dedicated data network on InfiniBand, Ethernet, or RoCE.

Server

In general, the servers of HPC systems are referred to as nodes. The number and configuration of the required nodes depends on workload, role, and budget. Although most nodes are used to run applications that process data, some nodes manage the storage and compute resources, or provide helper services such as remote access for data ingest and data exchange with peer sites.

As mentioned in the previous section, the nodes of an entry-level HPC system are typically off-the-shelf, rack-mounted servers composed of one or two CPUs and 256GB to 1TB memory (RAM). The reference architecture supports integrating varying node configurations, such as a mixture of x86_64 (Intel, AMD) or POWER (IBM, OpenPOWER) and optional accelerators such as GPUs or FPGA.

NetApp recommends using Linux as the operating system. For NetApp E-Series systems and IBM Spectrum Scale for supported Linux distributions, versions, and kernel levels, see the following support matrices:

- NetApp E-Series:
  - SANtricity software: Boost performance, maximize uptime
  - NetApp Interoperability Matrix (IMT)
- IBM Spectrum Scale:
  - Spectrum Scale Supported Operating Systems

Other operating systems can also be used, but in practice most customers use Linux.

Storage: NetApp E-Series storage systems

NetApp E-Series storage systems are designed to deliver the storage requirements of a broad range of workloads. High-performance file systems such as IBM Spectrum Scale and data-intensive workloads benefit from the ability of the E-Series systems to sustain high read and write throughput. Database-driven transactional applications benefit from the systems' high IOPS and low latency. Regardless of the application workload, NetApp E-Series systems are designed to support maximum performance efficiency.

NetApp E5700. Designed specifically for performance-intensive workloads environments, the NetApp E5700 hybrid flash system delivers more than one million sustained IOPS and response times in microseconds. Bandwidth-oriented workloads, such those that include parallel file systems and technical computing, also benefit from the ability of the E5700 array to provide up to 21GBps of throughput. The E5700 array is also the first hybrid 2U array to support multiple high-speed host interfaces, including 32Gb FC, 25Gb iSCSI, 100Gb InfiniBand, 12Gb SAS, 100Gb NVMe over InfiniBand, and 100Gb NVMe over RoCE. The hybrid design is built in a 2U or 4U enclosure and delivers the performance of more than two thousand 15,000-RPM drives while requiring less than 2% of the rack space, power, and cooling. With up to 98% reduction in space and power consumption, the E5700 hybrid array helps to significantly improve the overall efficiency of IT operations while continuing to meet performance requirements for business operations.

The E5700 system offers multiple form factors and drive technology options to meet your requirements. The ultra-dense 60-drive system shelf supports up to 600TB in just 4U and is optimal for environments with vast amounts of data and limited floor space. The 2U, 24-drive system shelf combines low power consumption and exceptional performance density with its cost-effective 2.5-inch drives. All shelves support E5700 controllers, or they can be used for expansion, helping you to optimize configurations to meet your performance, capacity, and cost requirements.

The E5700 hybrid array offers the industry's best price/performance ratio with a mix of media, including NL-SAS drives for capacity, SAS drives for cost-effective performance, and SAS SSDs for ultra-high performance. The E5700 array offers investment protection to meet future demands without forklift upgrades. It provides the ability to independently scale to 1.8PB of raw SSD capacity and 1.0M IOPS of performance, or up to 4.8PB of raw disk drive capacity and up to 21GBps of throughput performance.

NetApp E2800. The NetApp E2800 is a hybrid flash storage system with a low acquisition cost and even lower cost of ownership. Unlike other storage systems that add file or virtualization layers in the I/O data path, the E2800 hybrid arrays are purpose-built to optimize performance for mixed workloads. The E2800 system improves IOPS and throughput to help you extract value from your data and take action faster. The intuitive, on-box graphical interface simplifies configuration and maintenance while providing enterprise-level storage capabilities to deliver consistent performance, data integrity, reliability, and security.

The E2800 system optimizes price and performance to support any workload. Higher performance with SSDs enables the E2800 system to maximize storage density, requiring fewer disks to meet your performance objectives. The E2800 delivers over 300K sustained IOPS and supports a broad range of high-speed host interfaces, including 32Gb FC, 25Gb iSCSI, and 12Gb SAS, to protect your investment in storage networks.

SANtricity management software. The NetApp E-Series modular design and simple management tools make it easy to scale without adding management complexity. The on-box, browser-based SANtricity® System Manager GUI enables you to streamline deployment and start working with your data in less than 10 minutes. Simplicity doesn't preclude flexibility. NetApp SANtricity software offers a combination of comprehensive features and ease of use. You can use SANtricity System Manager to walk through workload-appropriate provisioning, or you can provision workloads on your own. NetApp Dynamic Disk Pools (DDP) technology dramatically simplifies RAID management by distributing data, parity, and spare

capacity across a pool of drives, with intelligent defaults, minimal decision making, and no stranded capacity after deletions. A single E-Series system can support both DDP and traditional RAID volumes if your workloads require it.

File system: IBM Spectrum Scale

IBM Spectrum Scale is an industry-leading, high-performance, parallel file system software. It powers the fastest supercomputers in the world. Large Spectrum Scale deployments store hundreds of petabytes of data in billions of files and directories. Entry-level HPC systems benefit from functional and performance enhancements that have been added to Spectrum Scale to meet the requirements of large HPC systems.

In this reference architecture, Spectrum Scale owns the virtual disks provided by the NetApp E-Series systems and makes their storage capacity available as file and object storage. The POSIX interface makes Spectrum Scale look like a local file system, which is very fast and very scalable.

A key ability is a single namespace (or data plane) so that each data source can add data to Spectrum Scale by using NFS, SMB, HDFS, S3 object, CSI, or a POSIX interface. This single data plane allows the data prep tools to access the data in place – no copying required. AI training can also access the data in place, as can the inference applications, all with no copying and all through industry-standard interfaces.

This reference architecture supports optional data access nodes, which expose the Spectrum Scale NFS, SMB, HDFS, and S3 object interfaces. This feature allows external users and applications to ingest and access data that is stored on the NetApp E-Series systems and share it with processes running inside the HPC system – again, with no copying required.

Spectrum Scale's many built-in interfaces allow traditional HPC applications, Hadoop and Spark, and containerized applications to access and share datasets, which are stored in a single Spectrum Scale file system. This ability enables and accelerates workflows such as the analysis of acquired data and AI pipelines that require multiple access methods and make it possible to consolidate the many applications on a shared infrastructure.

Another key strength is that Spectrum Scale enables data to be tiered automatically and transparently to and from more cost-effective storage, including NL-SAS disks, tape, and object. This reference architecture uses this ability to configure the NetApp E-Series system with SSDs and NL-SAS drives. It also integrates both media types into the same Spectrum Scale file system to meet the capacity, performance, and budget requirements described in the "Solution overview" section.

Job scheduler

Job schedulers simplify and optimize the management and use of shared compute resources by:

- Determining the resources of each compute node such as available CPUs, GPUs, FPGAs, and their types and available memory.
- Measuring the actual use of those resources and reporting them centrally.
- Maintaining one or more job queues with currently running and pending compute jobs.
- Dispatching pending compute jobs to one or more compute nodes, depending on job criteria such as priority, requested resources, and available resources.
- Enabling users to submit compute jobs to the job queue for batch and interactive processing.

Modern job schedulers provide a GUI and a single point of control for the definition, management, and monitoring of all resources by administrative users, and the submission and monitoring of jobs by end users. They integrate with frameworks such as Hadoop, Spark, and containers. For instance, this integration makes it possible to spawn and decommission Hadoop or Spark clusters on demand, and therefore to share the same set of compute nodes for a variety of workloads.

It is beneficial to have a job scheduler even for an entry-level HPC system. The job scheduler allows flexible sharing of compute resources between team members. It is also possible to dedicate some

compute nodes for interactive analysis during the day and to use all resources for batch processing during nights and weekends.

This reference architecture supports optionally using any job scheduler that is compatible with the operating system of the compute nodes. Examples of job schedulers include IBM Spectrum LSF, Univa Grid Engine, and SchedMD SLURM.

Network considerations

The nodes of an HPC system build a cluster computer. To accelerate the time-to-analysis results, they need to be connected through a high-speed, low-latency network. For instance, Message Passing Interface (MPI) applications require a low-latency connection between compute nodes to exchange application status and synchronize application execution. Compute nodes require low-latency and high-bandwidth access to the data to analyze acquired data and to run AI training jobs rapidly.

Experience in the field has proven that using existing networks can be problematic because they are not designed for high-throughput and low-latency I/O. Other activity on the shared network can cause degradation of the HPC system. Experience in the field has also proven that running the network over a shared infrastructure can be problematic. Features such as VLAN and quality of service on shared links must be configured carefully to support all cluster applications, including Spectrum Scale.

This reference architecture therefore includes a dedicated data network that must be established in addition to the existing networks to connect all nodes of the entry-level HPC system. This dedicated data network should not be connected to components outside the HPC environment, such as a data center network, a campus network, or the internet.

This reference architecture supports a variety of network technologies. The minimum requirement is 10GbE, and faster networks such as 25/40/100/200GbE and InfiniBand offer improved performance. InfiniBand with remote direct memory access (RDMA) provides the best performance. RDMA over Converged Ethernet (RoCE) offers better performance than plain Ethernet, but it must be configured carefully to provide the required performance and resilience.

Reference architecture

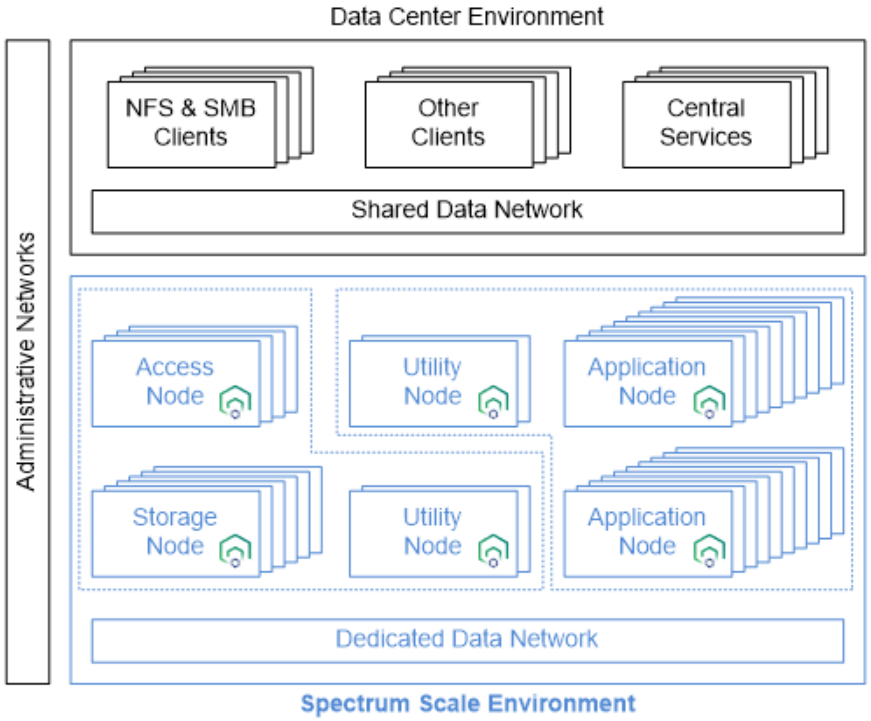
Spectrum Scale is the linchpin of the data that is processed in the HPC system. The Spectrum Scale deployment architecture is therefore a good thought model to structure the reference architecture. In this section, the structure of a Spectrum Scale deployment is first described and one example deployment is outlined to illustrate key architectural decisions. This section concludes with three example configurations for entry-level HPC systems.

Spectrum Scale environment

Spectrum Scale is delivered as software that is installed on the operating system of each node that needs to mount a Spectrum Scale file system. Spectrum Scale and the underlying hardware (servers, storage, network) build a Spectrum Scale environment (Figure 2). This environment is composed of one or more Spectrum Scale clusters, and each cluster is built on Spectrum Scale nodes. A typical Spectrum Scale environment includes:

- One application cluster
- One storage cluster
- One dedicated data network

Figure 2) Spectrum Scale environment.



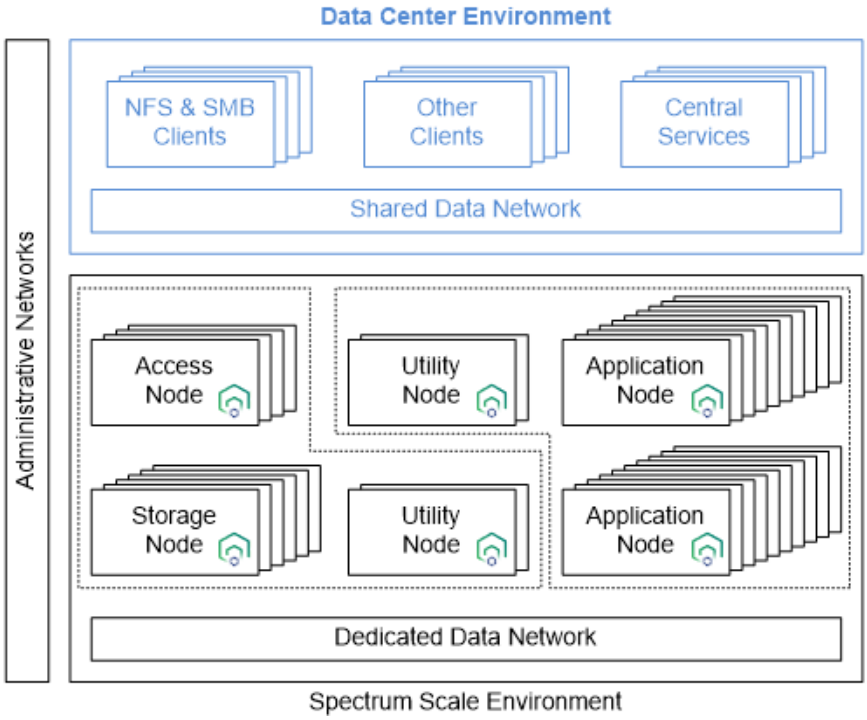
The dedicated data network connects all nodes of the Spectrum Scale environment. External users and applications access the environment through a shared data network such as a data center network, a campus network, or the internet. The boundary of a Spectrum Scale environment and the dedicated data network impacts its availability and performance. Therefore, it's necessary to carefully decide which nodes are part of the Spectrum Scale environment and which are not.

The storage cluster is composed of storage nodes, access nodes, and optional utility nodes, and the compute cluster is composed of application nodes and optional utility nodes. The node roles are defined as follows:

- **Application nodes** run applications to access and analyze data stored in one or more Spectrum Scale file systems. Most nodes of a Spectrum Scale environment are application nodes.
- **Storage nodes** provide access to underlying storage media and storage systems. In the reference architecture, all other nodes access the E-Series systems via storage nodes.
- **Data access nodes** provide access to Spectrum Scale file systems using protocols like NFS, SMB, HDFS, and S3 object. They enable applications and devices that cannot be attached to the dedicated data network to access and process data stored in the HPC system.
- **Utility nodes** are optional nodes that provide administration services, including general management and monitoring, and selected data management tasks such as backup, external tiering, and hybrid cloud workflows.

Components and services of the Spectrum Scale environment interact with components and services of the data center environment (Figure 3). Selected nodes of the Spectrum Scale environment are connected to the shared data network in addition to the dedicated data network to enable external data access and to connect to external services.

Figure 3) Data center environment.



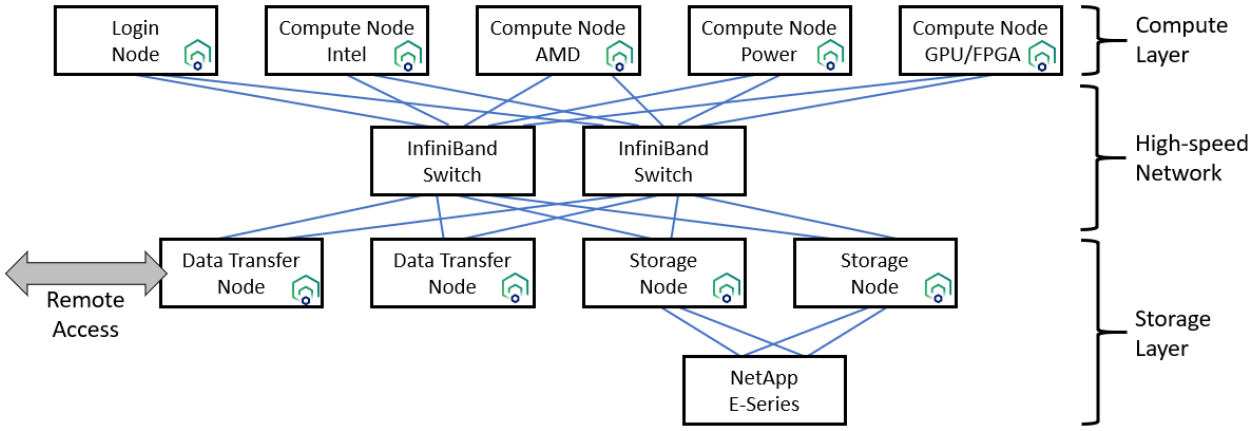
Relevant external components and services are described as follows:

- **NFS and SMB clients** give external users and applications access to data stored on Spectrum Scale file systems.
- **Other clients** give external users and applications access to data stored on Spectrum Scale file systems through additional access protocols such as HDFS, S3 object, IBM Aspera®, rsync, Secure Copy Protocol (SCP), and more. Other clients include servers for administrators and applications that manage components of the Spectrum Scale environment using GUI clients, REST API clients, SSH client, and more.
- **Central services** describes infrastructure services that are required for the whole solution, such as authentication and ID mapping (for example, AD, LDAP), time synchronization (for example, NTP), name resolution (for example, DNS), and more.

Example deployment

A Spectrum Scale environment, as described in the previous section, must be customized to create an entry-level HPC system that includes Spectrum Scale, E-Series systems, and an optional job scheduler. Figure 4 shows an example deployment. In contrast to Figure 2, the components of a Spectrum Scale environment are arranged differently to illustrate the cabling and layering.

Figure 4) Example deployment.



In this example, the dedicated data network is built on InfiniBand, because at this time InfiniBand is the predominant network technology for HPC systems. All nodes are connected through two independent InfiniBand switches to provide high availability. The section called Spectrum Scale Configuration describes configurations in which the compute nodes are connected via only one InfiniBand link, to reduce the cost of the overall solution.

The Spectrum Scale storage cluster builds the storage layer. The E-Series system is attached to two storage nodes for high availability. SAS or FC is the preferred technology for these connections. All other nodes are not connected to the E-Series system. They access the data on the E-Series system through one or more Spectrum Scale file systems and the storage nodes.

The two data transfer nodes enable secure access through protocols such as NFS, SMB, HDFS, and S3 object to ingest data and to share it across sites and institutions. Data transfer nodes therefore are connected to the InfiniBand network as well as to the external data center or campus network.

In HPC systems, application nodes are referred to as compute nodes. They execute batch jobs that are dispatched by the job scheduler. These nodes are less stable because end users might dispatch new experimental applications. The login is typically restricted to administrative users only, although compute nodes can be configured for optional user login to enable interactive applications for data exploration or data visualization. Spectrum Scale's proprietary access protocol is optimized for fast data access, so that I/O-intensive interactive applications can run on Spectrum Scale nodes much faster than on NFS or SMB clients.

Login nodes are utility nodes that allow users to log in to these nodes to compile applications or to submit compute jobs. Login nodes are stable nodes and therefore are good candidates to run additional infrastructure services for the HPC system.

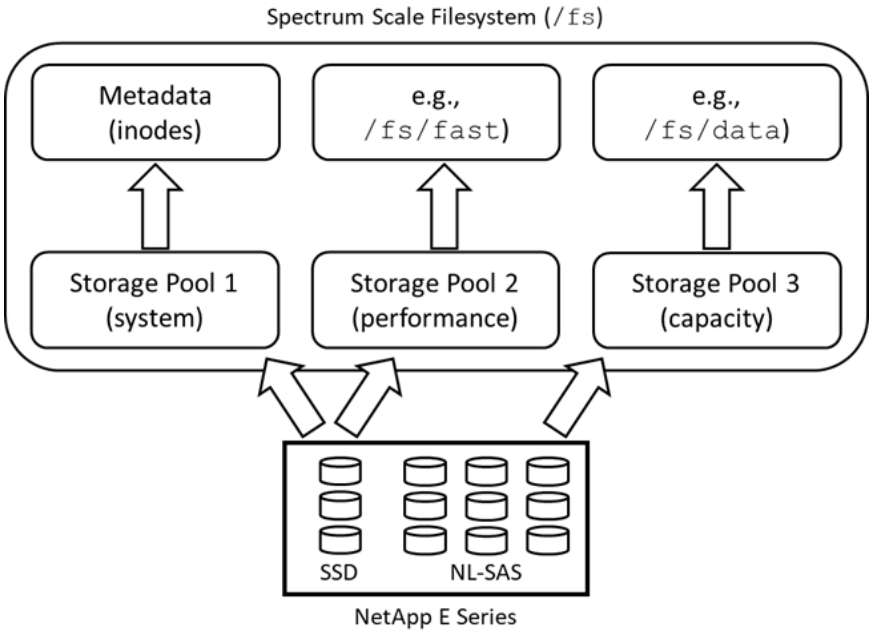
Depending on the configuration of the job scheduler, an HPC system might include additional management nodes to run all the services to dispatch and manage batch jobs. The login is restricted to administrative users only. Management nodes are the most stable nodes and are therefore good candidates to run additional infrastructure services for the HPC system.

Spectrum Scale configuration

Spectrum Scale can be configured with one or more Spectrum Scale file systems, independent of the number of E-Series storage systems. Spectrum Scale has built-in heuristics to stripe the load across all available resources (storage nodes, network links, and storage systems). Experience in the field shows that those built-in heuristics achieve good performance for many workloads, if all resources are provisioned to only one Spectrum Scale file system and therefore are shared across all users and

applications. To keep the configuration and operation simple, NetApp recommends configuring one Spectrum Scale file system that contains three Spectrum Scale storage pools (Figure 5).

Figure 5) Example file system configuration.



Having three Spectrum Scale storage pools allows the separation of SSDs for data and metadata and the integration of SSD and NL-SAS drives in a single file system. Spectrum Scale always stores the metadata (inodes) of files and directories in the system pool. Equipping the system pool with SSDs improves the user experience for interactive access when users run operating system commands such as `ls -l`.

Spectrum Scale has a built-in policy engine that makes it possible to store the contents of files in different storage pools. The reference architecture uses one storage pool for SSDs and one for NL-SAS drives. It is not obvious to users and applications in which storage pool the content of files is stored. Even within one directory, the content of some files can be stored on SSDs while the content of other files is stored on NL-SAS drives. Users and applications might notice a difference in performance when the files are accessed, but they don't see the difference in the name or path of files and directories.

The Spectrum Scale policy engine supports two kinds of policies: placement policies and migration policies. Placement policies direct Spectrum Scale which storage pool to use for new files, based on criteria such as file name, file path, and file owner. Migration policies direct Spectrum Scale to move the content of files between storage pools without changing file names or file paths, based on criteria such as file name, file path, file owner, and file access history.

The SQL-like syntax of the policy engine allows Spectrum Scale administrators to customize the policies for varying needs to optimize the usage of the more expensive SSDs. The following optimization strategies illustrate the power and the flexibility of the policy engine:

- Store all file data per default in the capacity pool. In the performance pool, store only files in the directory `/fast` or with the name `*.xyz`.
- Store all file data per default in the performance pool. Every night, or when a threshold is reached, move cold data from the performance pool to the capacity pool.
- Move the data of a specific dataset from the capacity pool to the performance pool before the job scheduler dispatches a compute task that accesses the specified dataset.

E-Series storage system configuration

E-Series systems provide the storage capacity for the entry-level HPC system. Experience in the field suggests configuring at least 10% of the storage capacity with SSD and the rest with NL-SAS drives. Having at least 10% of the storage capacity as SSDs ensures that there is enough headroom for the Spectrum Scale policy engine to effectively manage the data migration between the performance storage pool and the capacity storage pool.

NetApp recommends mixing SSDs and NL-SAS drives in the same E-Series system (hybrid configuration). Separation of SSDs and NL-SAS drives in different E-Series systems typically provides better performance. However, experience in the field has proven that entry-level HPC systems typically run mixed workloads for which hybrid E-Series configurations provide reasonable performance for a good price.

The SSDs of the E-Series system are used for the Spectrum Scale system storage pool and the Spectrum Scale performance pool. NetApp recommends using either RAID 5 (4+1) or DDP. RAID 5 provides better performance than DDP. DDP provides shorter rebuild times than RAID 5. The reference architecture supports both ways. In either way, configure 4 volumes per RAID 5 volume group or 8 volumes per DDP.

The NL-SAS drives of the E-Series system are used for the Spectrum Scale capacity storage pool. NetApp recommends using either RAID 6 (8+2) or DDP. RAID 6 provides better performance than DDP. DDP provides shorter rebuild times then RAID 6. This reference architecture supports both ways. Configure one volume per RAID 6 volume group or at least four volumes per DDP.

HPC system configuration

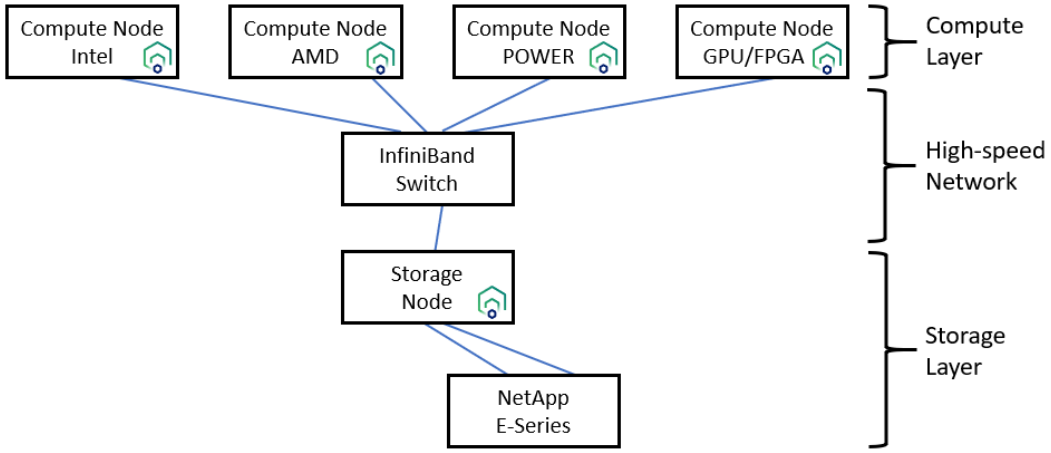
This reference architecture can be implemented on different hardware infrastructures. This section describes three system configurations for entry-level HPC systems that support all previously discussed requirements and features, but they vary in performance and resilience. Table 1 compares the three configurations.

Table 1) System configurations.

	Configuration 1 (cost minimized)	Configuration 2 (full HA)	Configuration 3 (cost optimized)
Compute nodes	2 to 10 servers	2 to 27 servers	2 to 46 servers
Storage nodes	1 server	2 servers	2 servers
User login nodes	1 server	2 servers	1 server
Management nodes	-	2 servers	2 servers
Data transfer nodes	-	2 servers	2 servers
Storage	1 NetApp E2800	2 NetApp E5700s	1 or 2 NetApp E5700s
Network	1 Mellanox SB7800 (EDR 100Gbps)	2 Mellanox QM8700s (HDR 200Gbps)	2 Mellanox QM8700s (HDR 200Gb/s)
Inter-Switch links			Variable
File system	IBM Spectrum Scale	IBM Spectrum Scale	IBM Spectrum Scale
Job scheduler (optional)	-	Variable	Variable

Configuration 1 is a very small configuration that is extremely cost optimized (Figure 6). Scaling and performance of the whole system is limited by the NetApp E2800 system. Therefore, the number of compute nodes was limited to 10. In many cases, InfiniBand EDR technology is sufficient to meet performance requirements. Users of this configuration must accept that storage and network failures can cause an outage of the whole HPC system. Choose configuration 1 if you have a very constrained budget and can accept the limited resilience.

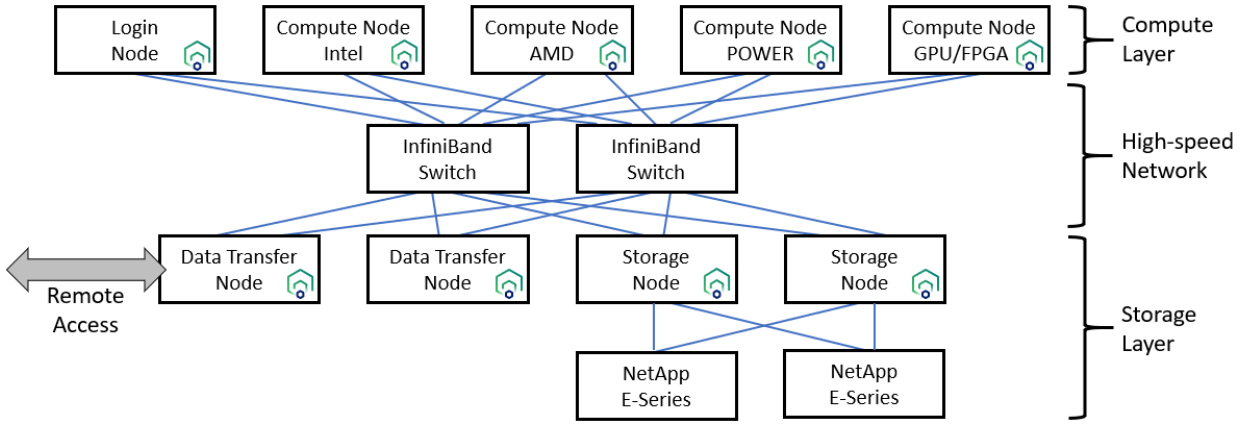
Figure 6) Layout of configuration 1 (cost minimized).



Configuration 2 is a small configuration that is optimized for high availability (Figure 7). The HPC system is resilient against hardware failures. NetApp recommends replicating the metadata (inodes) in Spectrum Scale across both E-Series systems. Customers can choose to replicate the content of all files or the content of selected files across both E-Series systems. All nodes are connected through dual-rail InfiniBand for full resilience against switch failures. Choose configuration 2 if you want very high resilience.

Scaling and performance of this configuration is limited by the number of switch ports. NetApp chose InfiniBand HDR technology to improve the overall system performance. A Mellanox QM8700 InfiniBand switch supports up to 40 HDR ports. HDR switches with more ports are available on the market, but they are too expensive for entry-level HPC systems. It is common practice to connect multiple 40-port switches to create InfiniBand networks that provide support for more servers and storage systems to create medium-size and large HPC systems, but this is outside the scope of this reference architecture.

Figure 7) Layout of configuration 2 (full HA).



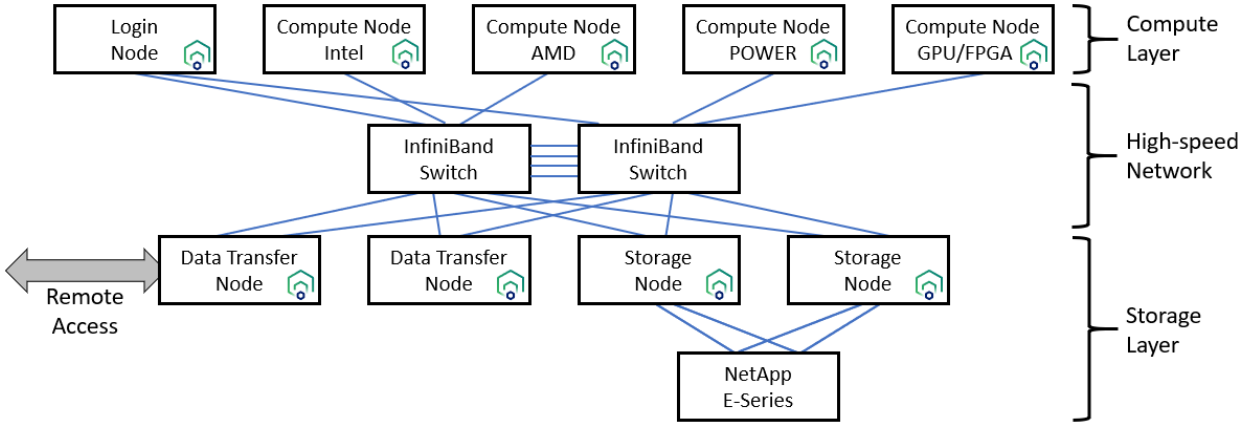
Configuration 3 is a small configuration that is cost optimized by allowing the addition of more servers to lower the resilience against selected component failures (Figure 8). Critical nodes such as storage nodes, management nodes, and data transfer nodes are connected through dual-rail InfiniBand for full resilience against switch failures. Compute nodes are connected through only one InfiniBand link to reduce the cost of the overall solution and to add more compute nodes compared to configuration 2. Failure of one switch might therefore cause the loss of 50% of the compute nodes. If two E-Series systems are available, NetApp recommends replicating the metadata (inodes) in Spectrum Scale across both E-Series systems

and optionally replicating the content of all files or the content of selected files. Choose configuration 3 if you want a balanced configuration with good resilience.

Scaling and performance of this configuration are limited by the number of switch ports. In contrast to configuration 2, compute nodes are connected through only one InfiniBand HDR link. Experience in the field has proven that a single InfiniBand link provides sufficient bandwidth for most workloads that run on entry-level HPC systems.

All nodes of a Spectrum Scale environment must be able to communicate through a low- latency network to maintain a consistent global state of all files and directories stored in Spectrum Scale file systems. Configuration 3 therefore includes four HDR Inter-Link Switch links to connect both InfiniBand switches for Spectrum Scale. Optional additional HDR Inter-Switch links can be added for parallel applications such as MPI.

Figure 8) Layout of configuration 3 (cost optimized).



Conclusion

This paper introduced a reference architecture for entry-level HPC systems. The architecture supports workloads such as entry-level HPC and HPC-like workloads including AI and analytics, consolidation of workstations, and replacement of NFS or SMB file service for data-intensive science and engineering. The combination of IBM Spectrum Scale and NetApp E-Series systems provides a flexible data layer that can be customized to meet varying performance, resilience, and budget requirements.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services, and applications to the right people—anytime, anywhere.

and IBM Aspera are registered trademarks of IBM Corporation. Other company and product names may be trademarks of their respective owners.

Where to find additional information

To learn more about the information that is described in this document, see the following websites:

- E-Series and SANtricity 11 Documentation Center
  https://docs.netapp.com/ess-11/index.jsp
- E-Series and SANtricity documentation resources
  https://www.netapp.com/documentation/eseries-santricity/
- NetApp Product Documentation
  https://docs.netapp.com

Version History

Version	Date	Document version history
Version 1.0	May 2021	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.

TR-XXXX-DESIGN-0321

NetApp