NetApp E-Series NVMe-oF with InfiniBand or RoCE

One of the first NVMe over Fabrics enterprise systems

Key Benefits

- Offers superior price/performance value, with sub-100-microsecond latency
- Runs over a high-speed 100Gbps InfiniBand or RDMA over Converged Ethernet (RoCE) transport layer and network connections, powered by Mellanox technology
- Supports mission-critical big data analytics workloads where data access response times are critical
- Supports persistent reservations that allow cluster software on the host side to access shared namespaces on the array

The Challenge

Increasingly, organizations are looking for ways to improve the speed and responsiveness of the applications that control their key operations. Because the performance of these applications is tightly linked to time to value and to stakeholder satisfaction, it is critical that they operate at the highest levels with maximum efficiency. Getting value and insights quickly and reliably from a range of mixed-workload environments can differentiate your organization and accelerate efficiency.

The Solution

The midrange NetApp® EF570 and E5700 storage systems turbocharge access to your data and increase its value. Requiring just 2U of rack space, the systems combine extreme IOPS, sub-100-microsecond response times, and up to 21GBps of bandwidth with leading, enterprise-proven availability features, including:

- Redundant components with automated failover
- Intuitive storage management with comprehensive tuning functions
- Advanced monitoring and diagnostics with proactive repair
- NVMe over Fabrics (NVMe-oF) support, providing faster performance and investment protection
- NetApp SANtricity® Cloud Connector, enabling backup to the cloud and data mobility across NetApp systems

Combined, these capabilities provide industry-leading price/performance, configuration flexibility, and simplicity in a compact package to help you make decisions quickly and securely.

Introduction to NVMe-oF

NVMe has become the industry-standard interface for Peripheral Component Interconnect Express (PCIe) SSDs, with a streamlined protocol and command set and fewer clock cycles per I/O. NVMe supports up to 64K queues and up to 64K commands per queue, which makes it more efficient than the existing SCSI-based protocols such as SAS and SATA.

The introduction of NVMe-oF makes NVMe more scalable while still benefiting from low latency and small overhead. The NVMexpress.org specifications outline support for NVMe-oF over Ethernet, over remote direct memory access (RDMA), and over Fibre Channel (FC).

■ NetApp®

Mellanox TECHNOLOGIES

NetApp Technology Partner

nvm EXPRESS

NetApp is a Promoter Member of the NVM Express group

```
┌─ User Space ──────────────────────────────────────────────────────────┐
│                                                                        │
│                        Applications                                    │
│                                                                        │
└───────────────────────────────────────────────────────────────────────┘
                                    │
                                    ▼
┌─ Kernel Space ────────────────────────────────────────────────────────┐
│                        File System(s)                                  │
│                                                                        │
│                        Block I/O Layer                                 │
│                                                                        │
│     SCSI Mid-Layer                      NVMe Driver                    │
│     (SCSI upper level drivers)              /dev/nvme*n*               │
│        /dev/sd*      Other drivers       Fabrics         PCI           │
│                                                                        │
│     SCSI Low-Level Drivers                                             │
│        HW Vendor-supplied                                              │
│                                                                        │
└───────────────────────────────────────────────────────────────────────┘

               Physical Devices
   Local SCSI Devices   SCSI HBAs   NVMe-oF HBAs   Local NVMe Devices
```
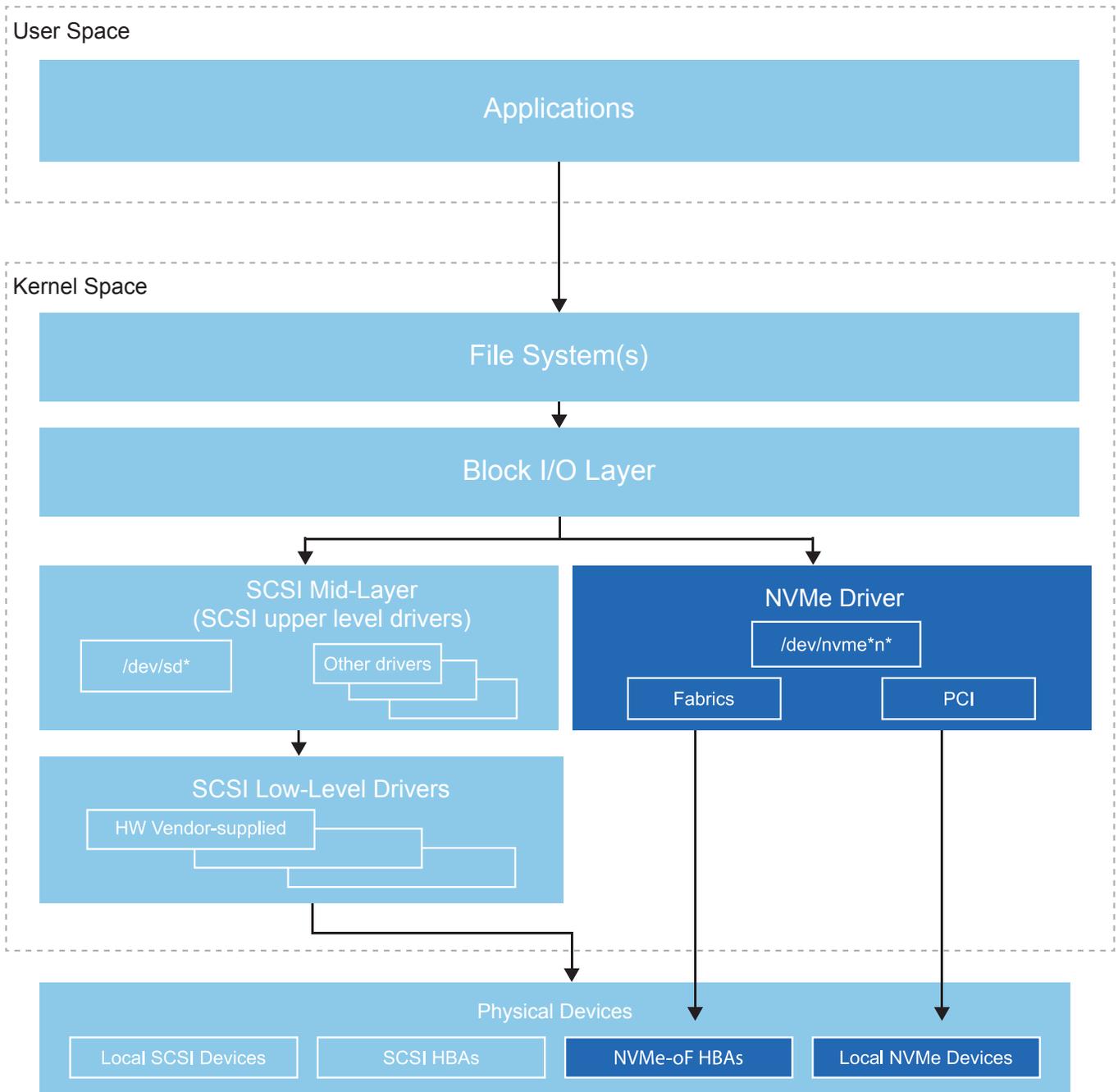
Figure 1) Linux OS driver stack.

NVMe-oF vs. SCSI

One of the advertised advantages of NVMe (and NVMe-oF), as compared with SCSI, is that it can support lower-latency I/O because the devices are faster and because the host OS driver stack (Figure 1) has inherent benefits. Because of this feature, I/O spends less total time getting from the application to the storage, thus reducing response times.

RDMA

RDMA is a technology for high-performance compute environments that enables a low-latency transfer of information between compute or storage nodes at the memory level. With Mellanox ConnectX InfiniBand Adapters, this function is offloaded to the network adapter to bypass the operating system network stack. Adapters therefore work directly with the application memory, eliminating the need to involve the CPU while providing a more efficient, faster way to send data.
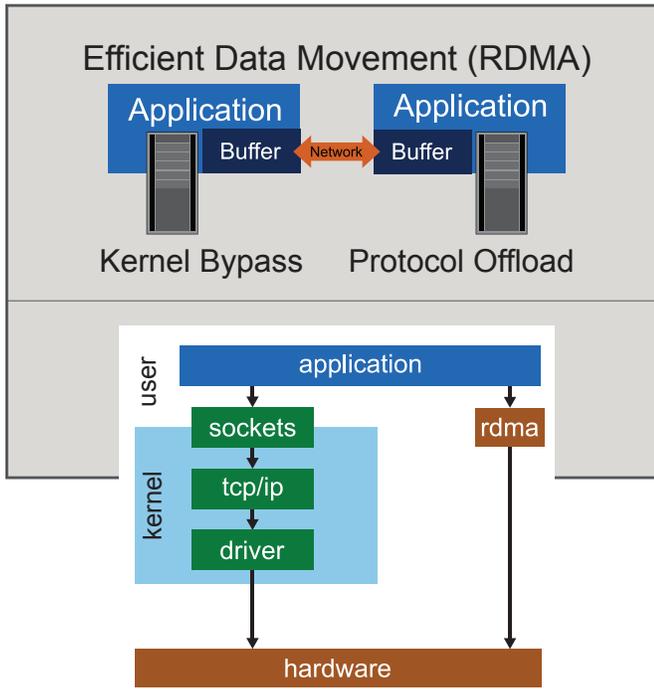
Figure 2) RDMA with Mellanox ConnectX™ InfiniBand Adapters.

RoCE

RoCE is a standard protocol that enables RDMA efficient data transfer over Ethernet networks. RoCE enables transport offload with hardware RDMA engine accleration and provides superior performance.

Achieve consistent microsecond response with 100Gb NVMe/RoCE host interface for EF570 and E5700 systems. Up to 1M IOPS, sub-100 microsecond latencies, and positioned for extremely low latency workloads such as Oracle databases, Microsoft SQL Server, and data analytics. Both the 100Gb NVMe/InfiniBand (shipping since 10/2017) and NVMe/RoCE host interface cards are powered by Mellanox technology.

NetApp E-Series

Most implementations on the market today focus mainly on just adding NVMe drives to the back-end storage while keeping the front end to the host SCSI based. NetApp E-Series, however, has taken a different approach. NVMe-oF is supported from the host to the front end of the EF570 and E5700 systems, while the back end is still SCSI based with the SAS drives. (See Figure 3.)

The NetApp E-Series implementation supports both NVMe over InfiniBand (NVMe/IB) and NVMe over RoCE (NVMe/RoCE) for several reasons:

- InfiniBand and RoCE have RDMA built into them.
- E-Series systems already have a long history of and experience in supporting other protocols over RDMA (SCSI based), such as iSCSI Extensions for RDMA (iSER) and the SCSI RDMA Protocol (SRP).
- The same hardware on EF570 and E5700 systems that runs iSER or SRP can run NVMe/IB or NVMe/RoCE (but not at the same time).
- The iSER, SRP, and NVMe/IB protocols can all coexist on the same fabric and even on the same InfiniBand Host Channel Adapter (HCA) port on the host side. Therefore, if you have an existing fabric that runs iSER and/or SRP, you can connect an EF570 or E5700 system that runs NVMe/IB to the same fabric.
- Both iSCSI and NVMe/RoCE can coexist on the same fabric on the host side.
- All the InfiniBand components in the fabric (NetApp EF570 or E5700, switches, HCAs, and cables/modules) can negotiate the speed down as needed (Enhanced Data Rate [EDR] of 100Gbps, Fourteen Data Rate [FDR] of 56Gbps, Quad Data Rate [QDR] of 40Gbps). This feature makes it easy to connect to legacy components with lower speeds.
- NetApp EF570 and E5700 systems support 100Gbps, 50Gbps, 40Gbps, 25Gbps, and 10 Gbps speeds for NVMe/RoCE.
- NetApp EF570 and E5700 systems support NVMe/RoCE v2, which is routable, and are backward compatible with RoCE v1.
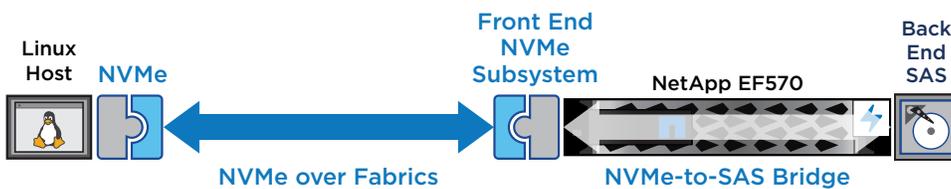

Figure 3) NVMe-oF front end on E-Series.

E-Series Interoperability with Third-Party Components

For a comprehensive list of supported components, see the NetApp E-Series Interoperability Matrix Tool (IMT). Following is an example:

- Operating systems: SUSE Linux Enterprise Server (SLES) 12 SP3
- Host adapters: Mellanox FDR and EDR HCAs for NVMe/IB, and for NVMe/RoCE, host adapters from other vendors such as Cavium (Marvel) and Broadcom
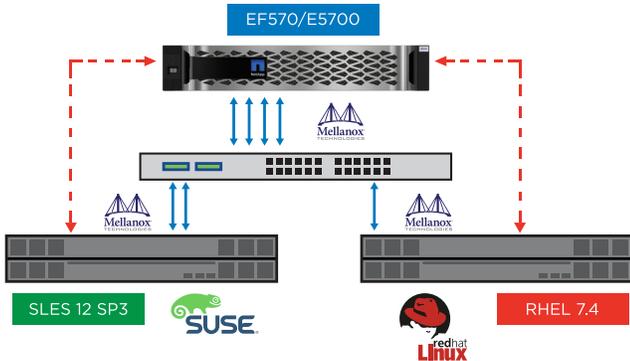- Switches: Mellanox, Cisco
- Topology: Fabric-attached and direct-attached

Figure 4) E-Series interoperability.

Coexistence Between NVMe-oF, iSER, and SRP

All three of these protocols can coexist on the same fabric and even on the same InfiniBand HCA port on the host side. Therefore, if your existing fabric runs iSER and/or SRP, you can connect an EF570 or E5700 system that runs NVMe-oF to the same fabric. Figure 5 shows an example.
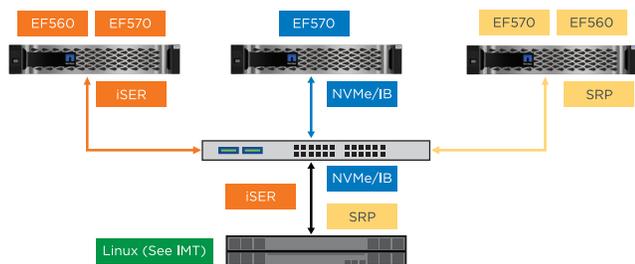

Figure 5) NVMe-oF, iSER and SRP coexistence.

Coexistence Between iSCSI and NVMe/RoCE

On the host side, iSCSI and NVMe/RoCE can run on the same network adapter (if it supports RDMA) at the same time, but they must be connected to different E-Series systems. Figure 6 shows an example.
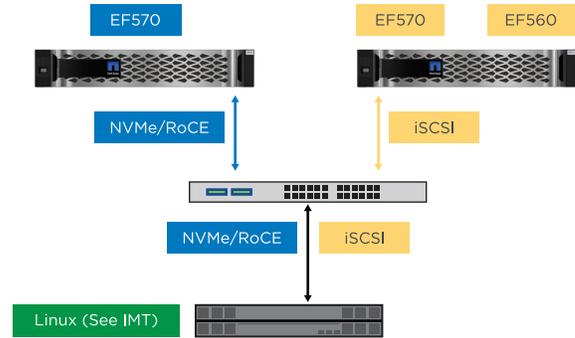

Figure 6) iSCSI and NVMe/RoCE coexistence.

About Mellanox

Mellanox is a leading supplier of end-to-end Ethernet and InfiniBand intelligent interconnect solutions and services for servers, storage, and hyper-converged infrastructure. Mellanox intelligent interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance. Mellanox offers a choice of high-performance solutions: network and multicore processors, network adapters, switches, cables, software and silicon, that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage, network security, telecom and financial services. **www.mellanox.com**

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organizations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation and optimize their operations. For more information, visit **www.netapp.com**. #DataDriven