Ethernet Storage Fabrics

Techie's Guide

MLNX-15-060437 | SEP 2020

Document History

MLNX-15-060437

Version	Date	Authors	Description of Change
2.0	September 22, 2020	Daureen	Rebranded template
1.0			Initial release

Table of Contents

Preface

The emergence of Ethernet as a storage networking fabric isn't new. We've seen it develop over the last 20 plus years. Starting in the early '90s with Network Attached Storage (NAS) and then evolving to add iSCSI Storage Area Networks (SANs). The influx of Ethernet-attached storage has consistently grown faster than the storage market itself, indicating that it is the fastest-growing storage fabric. Given the continued growth and adoption by emerging technologies, it seems realistic to anticipate that Ethernet storage will continue to dominate file and object storage and soon pass Fibre Channel (FC) in market share for block storage.

One thing is for sure, the digital data explosion isn't going away. It is invigorating the market for new and more efficient data storage solutions, which has opened the door for the emergence of Ethernet Storage Fabric™ (ESF). Why is Ethernet the future storage networking technology? First, Ethernet is ubiquitous and is the standard communication protocol used to facilitate worldwide communications, including the world-wide web and cloud. Similarly, every enterprise has standardized on Ethernet within their data centers due to its robustness, flexibility, and cost. Accordingly, enterprises have invested in Ethernet networking expertise, which in turn has driven a large pool of educated network engineers.

Powerful CPUs, large memory footprints, and the growing trend for server virtualization and containerization have driven the need for faster connectivity resulting in the introduction of 25/50/100 Gigabit Ethernet connections to servers and storage. The volume of data is growing in both structured and unstructured forms. Applications that generate or consume this data are being developed and deployed, both in centralized data centers and at the edge. All these trends call for a new storage infrastructure that is easy to expand with rapid data growth, fast to support new servers, storage and applications, agile to accommodate virtualized infrastructure, and efficient to operate at scale. Traditional storage, exemplified by complex, expensive, and proprietary FC SAN-based storage systems, cannot meet these requirements. As a result, modern data centers are breaking away from the "Big-Box" storage model and migrating to Ethernet-based storage.

Software-defined, scale-out storage, and hyper-converged infrastructure (HCI), have all standardized on Ethernet and are driving the need for higher speeds, which Ethernet answered with 25/50/100 G. Software-defined, scale-out storage and HCI each lay the foundation for today's cloud infrastructures–private, public, or hybrid–to achieve ultimate cost and operational efficiency while meeting ever-increasing demands for performance and capacity. However, adopting a cloud-like infrastructure, is only halfway to a data center transformation. The network fabric connecting these infrastructures also needs to be "modernized" before one can fully realize actual benefits, which is why you need an Ethernet Storage Fabric.

This eBook is designed to help you understand the essentials of an Ethernet Storage Fabric and how business benefits it provides as an underlining storage infrastructure. NVIDIA® Mellanox® end-to-end Ethernet portfolio can be combined to form a high performance, efficient, storage- optimized solution. Together they enable the ideal Ethernet Storage Fabric:

- Mellanox ConnectX® SmartNICs Ethernet adapters are RDMA-optimized, support Zero Touch RoCE plus storage, cloud, and security offloads (web link)
- Mellanox BlueField® SmartNIC Data Processing Unit (DPU) combines ConnectX adapters with advanced software and FPGA programmability (web link)
- Mellanox NVMe SNAP™ brings virtualized storage to bare-metal clouds and makes composable storage simple (web link)
- Mellanox Spectrum® Ethernet switches offer predictability, low-latency and high-bandwidth, are RoCE ready and provide zero packet loss for the best storage fabric (web link)
- Mellanox LinkX® cables provide an industry-low Bit Error Rate (BER), lower latency, and lower power for superior connectivity (web link)
- Mellanox Onyx® advanced Ethernet switch operating system optimizes ESFs by providing key automation, visibility, and management features that simplify storage fabric management (web link)
- Mellanox What Just Happened® (WJH) advanced telemetry provides real time visibility into network problems for root cause analysis (web link)

History of Networked Storage

Early computer storage devices were local hard disk drives, internal to the server and directly connected to a single computer through dedicated interface devices. As storage requirements changed, manufacturers shifted to directly attached storage (DAS) solutions, which were external but still connected to just one server. Soon these storage arrays were capable of being connected directly to more than one server which improved utilization but made resource management and maintenance more complex. Next, the concept of sharing stored data among independent networked computers introduced network-attached storage (NAS) and SANs. As NAS and SAN have co-evolved at an ever-increasing speed, at least within the data center they have virtually eliminated distance as a constraint to networked storage performance. The arrival of public, private, and hybrid clouds, object storage, and faster storage devices such as NVMe and 3D XPoint SSDs is driving the next evolution in networked storage—using Ethernet as a common infrastructure interconnect.

The evolution of storage has been driven by the changing ways in which we use and access data, by the exponential increase in the volume of data, and by the speed, flexibility, and cost efficiencies offered by an Ethernet Storage Fabric. This guide provides an in-depth look at data storage, answers the question, "why Ethernet is the future," and offers information on migrating to an ESF. Let's begin with a basic understanding of storage terminologies.

Protocols

Implementing storage is an important consideration for a data center, and that includes choosing which protocol(s) to use. There are many factors to consider, based on cost, performance, throughput, data sharing, and others. However, the market never stays idle for long, and there is a significant shakeup in Enterprise storage being driven by hyperscale and cloud deployment models. Let's look at the primary storage protocols and how they play within the data center.

SCSI/iSCSI

Let's start by discussing, SCSI (pronounced "skuzzy"). The Small Computer System Interface (SCSI) was developed by the American National Standards Institute (ANSI) as a set of parallel interface standards as a data access protocol in the early 1980. Its main use was in attaching printers, disk drives, scanners, and other peripherals to computers in its early years. SCSI was originally used to move data within a single server and is still today one of the dominant block-level access methods for data storage and SCSI did not require control mechanisms to handle data loss or have contention with

other network protocols. SCSI is also the foundation of Serial-Attached SCSI (SAS), iSCSI, and FC-SAN, all of which are commonly used to attach enterprise storage today.

ISCSI is a transport layer protocol that defines how SCSI packets are transported in the payload of IP packets. This allows the SCSI protocol to be extended outside the physical server and transported across Ethernet networks, making it possible to set up a shared-storage network where multiple servers and clients can access central storage resources as if the storage were locally connected. Initially, iSCSI storage systems were positioned as alternatives to the more expensive, yet higher-performing Fibre Channel-based storage arrays that handled the bulk of block storage tasks in enterprise data centers. iSCSI has been adopted rapidly in the cloud but has gained market share slowly over the years in the enterprise, mainly due to the relatively slow speeds of Ethernet infrastructure before 2015.

Fibre Channel/FCoE

Fibre Channel (FC) was designed to extend the SCSI protocol over longer distances by encapsulating SCSI data into FC frames on a dedicated and proprietary FC network. FCoE latter allowed the FC protocol (with encapsulated SCSI commands) to run on Layer 2 Ethernet in small environments such as a single rack of servers (usually one rack of Cisco UCS servers). Both FC and FCoE support only block storage traffic and cannot be used for compute, management, file storage, or object storage traffic. FC has declined in popularity due to its inflexibility, high pricing, and absence from cloud infrastructure. FCoE has nearly disappeared due to its limitations, with little to no interest in developing it on 25/50/100/200 GbE speeds.

NFS/SMB

NFS (Network File System) is a file-based storage protocol. NFS is traditionally used in Linux and Unix environments. NFS is also a widely used protocol for VMware environments and can offer several benefits for virtual machine storage and HPC or technical computing storage. File-based storage relies on an underlying file system such as FAT, HFS, VMFS, NTFS, or others. Block-based storage differs in which no underlying file system is required (though sometimes a file system is run on top of block storage). File-level storage is an excellent medium for some applications, especially when data sharing is required. NFS deployments usually work with most operating systems and can support some databases, VMware, and many engineering or big data applications.

SMB or Small Message Block is a file-based storage system based on CIFS (Common Internet File System) and typically used in Microsoft Windows environments for file sharing. Windows-based file shares rely on SMB as a transfer protocol at the file level and usually rely on an underlying file system such as FAT32 or NTFS in Windows environments. File-level storage is an excellent medium for some applications and many Windows-based applications—including SQL Server, SharePoint, and Exchange—can use SMB-based file storage as an alternative to block storage. However, for Windows applications needing block storage, SMB is not an option. For more information on SMB, see the Microsoft Storage Spaces Direct (S2D) section below.

Storage Systems

A storage system includes the storage controller and houses physical drives such as a hard disk or solid-state drives. It allows for the import and export of data to physical servers. A storage system is one of the critical components of a computer system or network and can be classified into several forms. To better understand the advantages of an ESF, it's essential to understand the different forms of storage systems, which are listed below.

JBOD

JBOD or "just a bunch of disks" commonly refers to a collection of hard disk drives typically contained in a single drive enclosure. JBOD is an alternative to using a storage controller-based RAID configuration. Rather than configuring RAID within the storage, a server using JBOD spans the disks or treats them as independent disks, and if redundancy (such as RAID or mirroring) is needed, it is done at the operating system or application level. Spanning configurations often use a technique called concatenation to combine the capacity of all the disks into a single, large logical disk.

While JBODs were popular in the early 2000s, they declined in popularity compared to RAIDs, though they enjoyed a small resurgence when Microsoft Exchange recommended them for a cost-effective architecture when used in conjunction with Database Availability Group (DAG) in Microsoft Exchange 2010 and beyond with Windows's failover cluster manager in Server 2012. Today, JBODs continue their resurgence in some Big Data applications as data sets become larger and more intricate. Several Hadoop distributions, including Cloudera's popular CDH, recommend using JBOD configurations when building Hadoop clusters where Hadoop is highly fault-tolerant at the application level due to the Hadoop replication facility.

NAS

A NAS or networked attached storage is a storage device that allows for a file-level storage access over an Ethernet network, usually via the NFS or SMB protocols. This allows for the storage and retrieval of data from a network-connected storage device that can be centrally located. NAS systems are flexible and most can be easily scaled-out as additional storage is required. With a NAS system, data is sharable and continually accessible to multiple servers or users include those at remote locations through VPNs, sort of like having a private cloud storage repository in the office. They are generally fast, affordable, and provide many of the benefits of a public cloud on-site, giving the administrator complete control. NAS is growing in popularity for certain types of big data, artificial intelligence, high-performance computing, and media and entertainment applications.

SAN

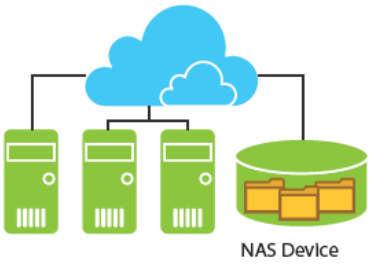
SAN or storage area networks are computer networks that provide access to consolidated, block-level storage for data. SANs were developed to provide enhanced access to storage devices, designed to enable centralization of storage resources and overcome direct attach storage or JBOD limitations. Improved access allows SANs to effectively, uncouple storage from the server and to pool resources onto a network where it can be easily shared, provisioned, protected, and managed, without the utilization and scaling problems associated with DAS.

While most storage networks utilize the SCSI protocol for communication, there are many mapping layers to other protocols that can be used to network. The most common are listed below:

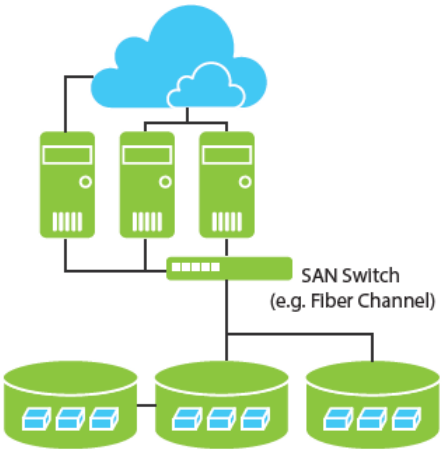
- Fibre Channel Protocol (FCP), the most prominent
- Fibre Channel over Ethernet (FCoE), the mapping of FC over with RDMA Ethernet, which has been fading in popularity
- ESCON over Fibre Channel (FICON), used by mainframe computers to attach FC storage
- iSCSI, which maps SCSI over TCP/IP
- iSCSI Extensions for RDMA (iSER), mapping of iSCSI over Ethernet or InfiniBand
- SCSI RDMA Protocol (SRP), running SCSI on top of InfiniBand

SANs were originally restricted to high-end solutions with Fibre Channel being deployed as the most widely adopted SAN interconnect technology in enterprises, iSCSI the most popular interconnect in the cloud, and InfiniBand the most popular interconnect for HPC.

Network Attached Storage



Storage Area Network



VMware VSA

VMware supports a hyperconverged infrastructure (HCI) solution called VSAN. It distributes applications, virtual machines, and management across a cluster of servers that each contain compute, storage and networking. VSAN clusters can be deployed quickly and expanded easily, allowing enterprise customers to rapidly build a private cloud infrastructure that is easy to scale. VMware VSAN—like most HCI systems—generates large amounts of east-west traffic between nodes for replication, migration, and management, so it benefits from the high bandwidth and low latency of an ESF. In addition, VSAN has previewed RDMA (RoCE) support since 2018 and is expected to have full support for RoCE in 2020, offering additional performance advantages for using an ESF.

Fibre Channel SAN

The Fibre Channel standard defined a high-speed data transfer mechanism that mapped the SCSI protocol to the underlying FC connections. This allow for connecting workstations, servers, mainframes, and supercomputers to storage devices.

The First Storage Fabric

Fibre Channel was the first storage technology to use a switched fabric for interconnecting data storage devices to servers. A Fibre Channel switched network or fabric provides a high-performance: any-to-any interconnect for server-to-server or server-to-storage traffic. A switched fabric network allows for a wide variety of connected devices and spreads network traffic out across multiple physical links. In doing so, it's able to yield higher total throughput. However, Fibre Channel switches are one of the costliest hardware devices in the world of networking or storage.

Why it Made Sense 20 Years Ago

Data storage requirements can quickly outgrow the limitations of a traditional SCSI bus. SCSI offers a highly reliable, high-performance data channel between a system and locally connected storage but is limited both in the scalability of 16 devices and cable length of up to 6 meters (for the original parallel SCIS standard, faster speeds permitted for shorter lengths).

Storage Area Networks provide a way to manage an enterprise data storage system. One of the key concerns for an IT department is the need for scalability. Additional servers and storage devices can be added to a SAN without bringing any of the existing devices offline. Another key concern in the modern business environment is data availability. To address this need, SAN-attached devices implement RAID and can be used as the shared storage for failover clusters. In addition, a SAN can be designed to offer redundant paths between attached devices, remote data replication, deduplication, compression, snapshots, and other storage management features. In a SAN environment, it is possible for multiple servers running different operating systems to be attached to the same storage devices.

When first implemented, Fibre Channel ran at faster speeds than Ethernet and offered a lossless network that was dedicated only to storage. At the time—25 years ago—Ethernet was running at 100 Mb/s or 1 Gb/s speeds and lacked strong solutions for flow control, lossless, QoS or congestion management. FC-SAN started at 1 Gb/s and then grew to 2, 4, and eventually 8 Gb/s speeds while Ethernet largely remained at 1 Gb/s.

Why Fibre Channel No Longer Makes Sense

Fibre Channel is 3X expensive as Ethernet with 1/3 the performance.

Times are changing, no longer do big enterprises generate and store most of the world's data inside their data centers. Cloud computing and the Internet of Things introduce a major shift, from block storage to file and object storage. Hard disk drives are being replaced with more performant flash storage. Companies that are embracing new technologies such as Big Data, artificial intelligence, and machine learning are gaining market share and are disrupting older legacy businesses that have been

slower in adapting and transforming. Over the past two decades, data has become one of the top assets for corporations, and mobile, social, and Internet-of-Things (IoT) have become significant data producers.

The volume of data, in both structured and unstructured forms, is multiplying rapidly. Applications that generate or consume this data are being developed and deployed across geographic regions. All these are calling for a storage infrastructure that is fast to expand with rapid data growth, agile to accommodate application performance requirements and virtualized infrastructure, and efficient to operate at scale. Analysts predict that file and object storage revenues will grow at 24% per year, much faster than block storage revenues. This is because file and object storage are better-suited for the typical storage from the new world. File and object storage run on Ethernet (and InfiniBand in niche environments) but not on Fibre Channel.

Figure 1. Ethernet & Fibre Channel Switch Port Shipment

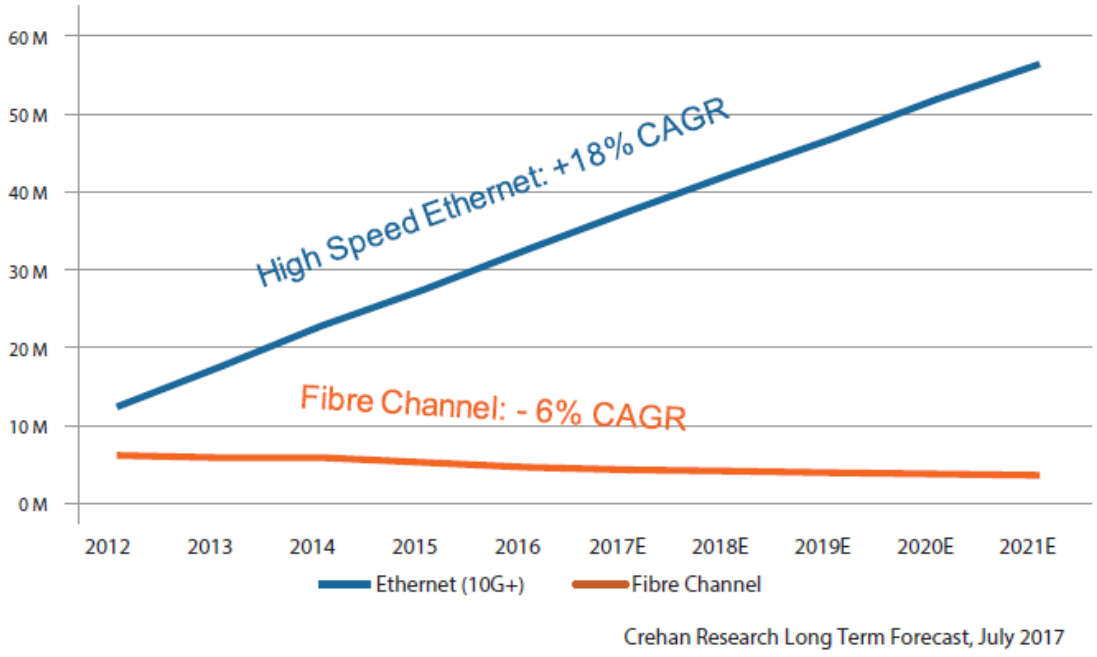
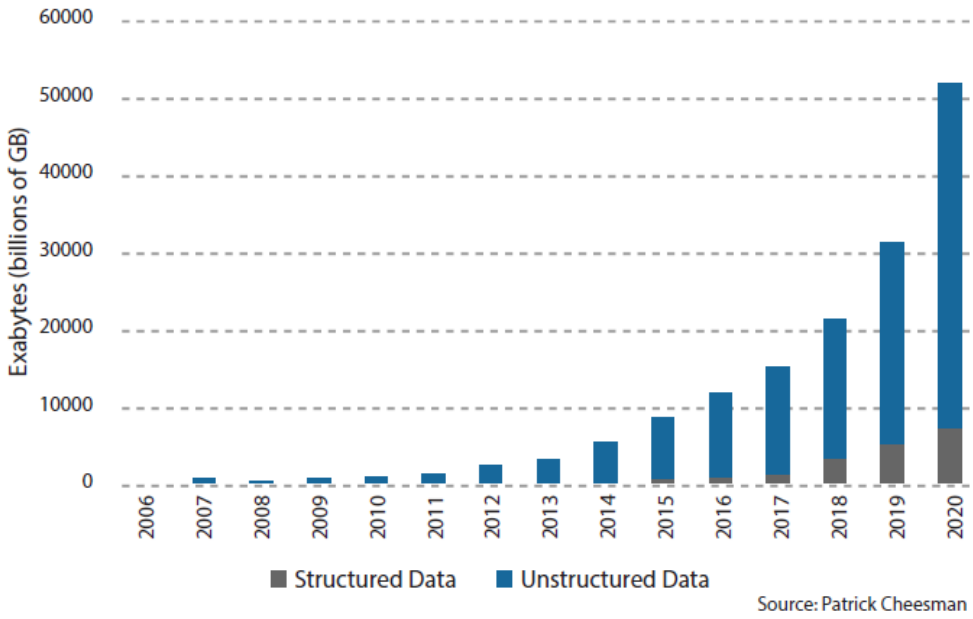


Figure 2. Unstructured data accounts for more than 90% of all data being created



Fibre Channel, exemplified by complex, and expensive SAN-based storage systems at 3 times the cost of Ethernet and 1/3 the performance, cannot meet modern data center requirements which is why you won't find FC in the cloud. The proprietary technology that loses ports to ISLs, involves LUN masking and zoning, all require FC expertise. Even with an experienced SAN administrator, congestion can occur if slower device ports converge with higher port speeds or switch bandwidth is not enough due to oversubscription. As a result, today's data centers are breaking away from the "Big-Box" storage model, and migrating to scale-out, software-defined storage (SDS) and hyperconverged infrastructure (HCI), where it's fast to deploy, elastic in scale, and flexible in the provisioning.

Modern Ethernet supports speeds of 100 Gb/s and 200 Gb/s per link, with latencies of just several microseconds, and combined with the hardware-accelerated RDMA, Ethernet allows iSER, and NVMe-oF protocols that are perfect for supporting maximum performance on flash or next-gen persistent memory (PM) solid-state storage. With fast storage, your physical network and your network protocol must have high bandwidth and low latency otherwise, you're wasting much of the value of flash.

Today's Ethernet runs at 25, 40, 50, 100 or 200 Gb/s speeds, is lossless, and no longer dependent on TCP alone. Ethernet easily supports multiple storage protocols—block, file, object—simultaneously, and allows application, management, and storage traffic to all share the same network, using traffic prioritization and QoS. Meanwhile, Fibre Channel is still deploying 16 Gb/s (actually 14 Gb/s) technology and only slowly starting to ramp up sales of 32 Gb/s-which still only support block storage traffic. Other storage (and other non-storage) traffic are already traversing and must remain on an Ethernet network.

Introduction to Ethernet Storage Fabric

An Ethernet Storage Fabric, or ESF in short, is the fastest and most efficient way to network storage. It leverages the speed, flexibility, and cost efficiencies of Ethernet with the best switching hardware and software. It comes packaged in ideal form factors to provide performance, scalability, intelligence, high availability, and simplified management for storage.

Figure 3. Ethernet Storage Fabric

An ESF is optimized for scale-out storage and HCI environments because it is designed to handle bursty storage traffic, move data with low latencies, provide predictable performance, and allow for simplified scale-out storage architectures with storage-aware services. These are all crucial attributes for today's business-critical storage environments. In particular, the switches must support new, faster speeds, including 25, 50, and 100 GbE. The ideal ESF fabric switch has an intelligent buffer design that ensures fast, fair, consistent networking performance using any combination of ports, port speeds, and packet size.

Additional ESF attributes include support for not just block and file storage, but also for object-based storage and hyperconverged infrastructure, along with storage connectivity for the newest NVMe over Fabric (NVMe-oF) arrays. Additionally, an ESF must provide support for storage offloads, such as RDMA, to free CPU resources and increase performance. Not only is an ESF specifically optimized for storage, but it also provides better performance and value than traditional enterprise storage networks More details on the distinction between an Ethernet Storage Fabric and storage deployed on an Ethernet network can be found in the Ethernet Storage Fabric – A Deep Dive section below.

Most Storage is Networked with Ethernet

In this age of rapidly expanding data, much of what is being created is from personal high-resolution photos and video files. These files are generated on personal devices; many are uploaded to social media and shared with friends and family. The only way this is possible is if the photo resides in the cloud. The cloud consists entirely of Ethernet storage. Even within the enterprise, rich media is stored on file or object storage, which is also almost entirely on Ethernet.

Today's virtual server environments are mostly leveraging Ethernet storage. With estimates of nearly 50 percent of Enterprise servers being virtualized today, there is plenty of growth ahead as most analysts agree that virtual server penetration will eventually reach 100 percent. As we head towards 100 percent, this will drive further adoption of Ethernet storage.

These providers rely on advanced technologies to services research needs, enterprise customers, and common mom and pop businesses. Service providers rely heavily on Ethernet storage. Likewise, one of the fastest-growing storage trends for enterprises is HCI, with its cloud-like appearance, which follows similar Ethernet store trends. With all the latest technology advances relying almost solely on Ethernet storage, why would anyone use older and less-flexible storage fabric technologies?

Ethernet is truly a ubiquitous technology in IT today. Ethernet storage has continued to expand from a few niche environments in the past to a significant percentage of attached storage today. Tomorrow's emerging technologies will unavoidably continue the trend of driving Ethernet storage as a preferred storage interconnect. Ethernet-connected storage is inexpensive and more versatile, but that's just the tip of its many benefits. More benefits will be covered, along with more background information within the next sections.

Key Data Center Disruptors

In the last few years, enterprises have been getting hungrier for higher data center performance. This has been answered by innovation in hardware, including more powerful processors capable of multithreading, higher core counts, and a shift to the PCI Express (PCIe) Gen 4 specification to offer higher bandwidth and more capacity. Enterprises are also starting to realize the

performance and latency benefits offered by new storage technologies featuring next-generation SSDs, high-speed NAND flash, and the new NVMe storage protocol. Faster computers and storage require faster, smarter networks to handle the volume of data, and innovators are turning to network infrastructure to provide the solution. The arrival of affordable, faster Ethernet with speeds from 25 to 200 Gb is enables high throughput with low latency. Networking storage over high-speed Ethernet with RDMA technologies (RoCE) enables networked storage to display local performance.

Compute, data and storage are moving to the cloud. Software-defined, scale-out storage and HCI each lay the foundation for today's cloud infrastructures – private, public, or hybrid – to achieve ultimate cost and operational efficiency while meeting ever-increasing demands for performance and capacity. However, by adopting a cloud-like infrastructure, you are only halfway through a data center transformation. The network fabric connecting these infrastructures also needs to be "modernized."

Automation fueled the Industrial Revolution and is driving a new wave of productivity within IT as it offloads manual tasks that drain productivity and agility. It's now possible to deliver scalable IT resources on demand. The results are shorter time to deployment, lower costs, increased agility, and a reduction in errors. However, data centers must be equipped with a flexible infrastructure to take advantage of automation.

Finally, there is this catch-all category of big data, which includes data analytics, artificial intelligence, and machine learning. They all share large files and big compute clusters to process or analyze data. Big data can be found in nearly every industry, including media and entertainment, oil and gas exploration, semiconductor design, automotive simulations, pharmaceutical research, finance, telco, and retail. On the data side, files can be multiple terabytes in size, and datasets can exceed a petabyte. On the analysis side, clusters can reach hundreds or thousands of nodes, processing millions of operations per second. As these clusters grow, flash storage is being used more often.

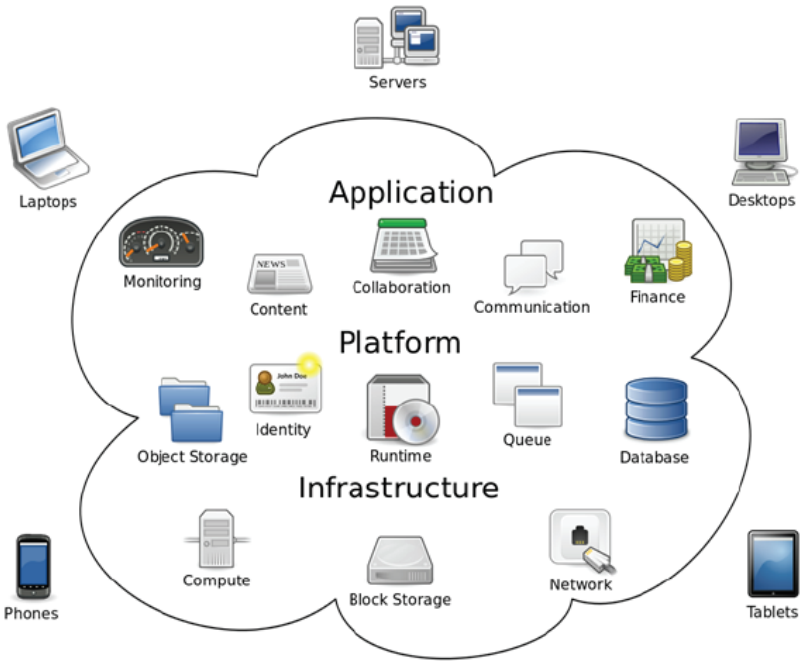
Sophisticated companies look for new technologies and embrace the shift toward cloud computing because it enhances data center efficiency. They welcome the introduction of software-defined models to automate business processes and adopt analytic software to discover exponentially more insights from their existing data. While each is a "key data center disruptor" – each brings with them new business models and cost advantages to the data center. Nevertheless, Ethernet is a common bond they all share to be effective. Ethernet allows for one network technology that can deliver

speed, flexibility, and cost efficiencies along with providing the foundation for the fastest and most efficient way to connect all types of servers and storage.

The Arrival of Cloud Computing

Perhaps the most innovative shift in computing in the last decade has been the introduction of cloud computing. The idea of providing a scalable method of delivering IT resources the same way utility companies deliver natural gas and electricity has revolutionized IT. Cloud computing packages system resources, such as computation, storage, and applications, then delivers them as a metered resource to the customer. Customers are offered on-demand computing and pay for only what is needed and used rather than a flat rate. The utility model maximizes the efficient use of resources and minimizes associated costs. This model has the advantage of a low or no initial cost to acquire computer resources; instead, resources are essentially rented.

Figure 4. Cloud Computing



To make this possible, several newly introduced technologies are engrained in the deployment of cloud computing. Below are some of the technologies required to lay the foundation for today's cloud infrastructures.

Commodity Servers

A commodity server is an inexpensive commercial off-the-shelf (COTS) computer, which usually has limited optional components. They can be easily purchased to help keep the cost down, which allows them to be disposable and replaced rather than repaired. Few general business computing requirements cannot be met with commodity servers, which is why cloud service providers utilize them.

Service providers also prefer to use many low-cost servers in parallel, with virtualization, then to use fewer high-performance, high-cost specialized servers. This creates another advantage as it increases reliability and resiliency as a larger number of discrete systems running in a cluster reduces the impact of any one server's failure.

Hyperconverged Infrastructure

Hyperconverged infrastructure (HCI) use software to combine compute, storage, and networking seamlessly to run on commodity, industry-standard x86 servers. These systems run virtualized or containerized workloads with a distributed architecture to take advantage of clustering multiple systems within data centers and spreading clusters across multiple sites. The cluster forms a shared resource pool that enables high availability, workload mobility, and efficient scaling of performance and capacity on demand.

Typically, HCI is managed through a single interface that allows users to define policy and manage activities at the virtual machine (VM) or container level which offers significant results including lower capital expenses (CAPEX) as a result of lower infrastructure costs, lower operating expenses (OPEX), and faster return on investment when launching new workloads.

One of the hottest areas of IT infrastructure today is HCI. It's growing quickly because it simplifies the rapid deployment of virtualized business applications and private or hybrid cloud. Most vendors like VMware, Microsoft, Nutanix, Dell, HPE, Lenovo, and EMC all offer HCI solution along with a handful of rapidly emerging startups such as Apeiron, Cohesity, Pivot3, and Qumulo. Because the network must carry both compute and storage traffic and be fast, efficient, and fast to deploy, it's always deployed on Ethernet and never Fibre Channel.

Software-Defined Storage

Software-defined storage (SDS) is software for policy-based provisioning and management of data that operates independently of the underlying hardware. With SDS, it is possible to install the software on commodify servers with the customer's choice of hard drives or SSDs, allowing the hardware to be tailored on the fly to specific workloads. SDS allows for storage hardware to be clustered for a specific purpose so multiple servers can operate as a single system. For example, one server cluster can be configured to hold user directories and NFS/CIFS folders, while another can be configured for block storage for database and yet another cluster is set up for photos and multimedia. Some SDS solutions can be used to consolidate and deliver more than a petabyte of data storage with a configuration time of fewer than 30 minutes.

Scale-out Storage

The emergence of software-defined storage has led to an explosion of scale-out storage solutions. Scale-out storage allows for a single logical storage system to be expanded through the addition of new hardware that is added and configured as required. When a scale-out system reaches its storage limit, additional arrays can be added to simply and non-disruptively expand the system capacity. Scale-out storage can also be used to add additional performance by adding new arrays that offer more storage controllers to connect to the added capacity.

Before scale-out storage, storage purchases would have to be much larger than current needs to account for future expansion, and most of the disk space sat idle for long periods of time, adding upfront cost to purchasing storage systems. If requirements turned out to be less than predicted, it was a waste of capital expenditure, and if multiple systems were deployed, they each needed to be managed separately. With a scale-out architecture, the investment can be smaller upfront, and when storage requirements increase, new arrays can be added as needed, virtually without limits, and all managed as one large system. Today, scale-out systems are available for just about any use case, from all-flash, to massively scalable object stores, and even for backups through integrated backup/storage appliances.

While Fibre Channel SANs promise the ability to scale capacity, it turns out that scale comes with an additional expense. Within FC SANs, at least two separate networks are required. There's the Ethernet network that interconnects servers to each other and a Fibre Channel network connecting the block storage to the clients. In addition, most Fibre Channel storage can only scale to a few controllers (typically 2, 4, 6 or 8), increasing the management burden for very large deployments.

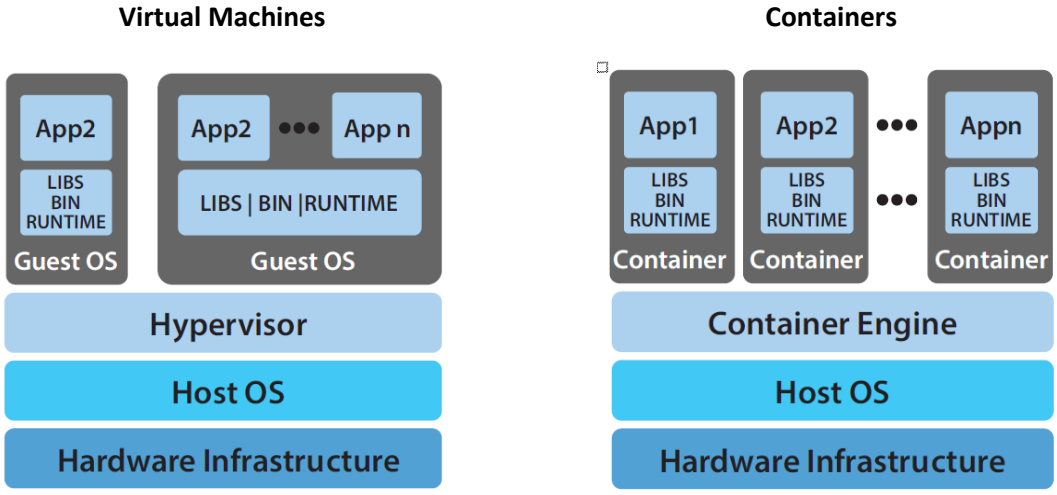
The simple way around this is to switch over to an ESF architecture. Instead of having two separate networks to buy and maintain, organizations simply connect to an ESF utilizing high throughput Ethernet networking technology, which can be deployed as one converged network or as separate storage and compute networks that use the same Ethernet technology. With an ESF architecture, capacity and performance can be quickly scaled without suffering any degradation in performance.

Virtualization

Server virtualization is a process that abstracts a physical server and is used to creates multiple virtual instances that run on the server. The virtual environments are referred to as virtual machines (VM), which enable the creation and management of many virtual instances of any operating system on the virtualization platform. Before server virtualization became mainstream, each operating system required a separate physical server with CPU, disk, memory, and other associated hardware to host the operating system. As servers have become more powerful, this approach has proven very wasteful on hardware resources, which typically ran at about 15% utilization efficiency. In typical server virtualization, many virtual machines are associated with each physical server, and a hypervisor is used to share the host hardware with each virtual machine, enabling organizations to increase the efficiency of each server significantly and reduce their data center hardware footprint.

Virtual Machines

A virtual machine (VM) is a software emulation program that is bootable on a physical machine and behaves like an actual computer. Much like any other program, it is accessed through a window, giving the end-user the same experience on a virtual machine as if they were active on a host system itself. Multiple virtual machines can run simultaneously on the same physical computer utilizing a hypervisor to manage them. Each virtual machine is designated a CPU, memory, hard drives, network interfaces, and other devices that are mapped to the actual hardware on the physical machine-reducing the need for physical hardware, the associated maintenance costs, and the power and cooling demands of the data center. The virtual machine has its own operating system and devices that run independently from other VMs operating on the same physical system. An isolated environment is ideal for conducting tests, accessing virus-infected data, creating operating system backups, and for running software or applications.



Configuring storage in virtualized environments is easy when using Ethernet attached storage as it only requires the IP address (or FQDN) of the target, plus the mount point. Datastores appear immediately after the host has been granted access. Fibre Channel is more confusing because it involves zoning at the FC switch level and LUN masking at the storage array level. Only after the storage target has been discovered through a scan of the SAN, LUNs are available for datastores. Fibre Channel also proves harder to troubleshoot than other protocols. Another difficulty of Fibre Channel is that it commonly only runs at 16 Gb/s, which is slower than typical Ethernet networks. (32 G FC is expensive, uncommon today, and throttled down to run at 16 Gb/s in vSphere 5.5.) FC-SAN also requires dedicated HBA in each host machine, FC switch ports, and an FC-capable storage array, which makes an FC implementation more expensive along with the additional management overhead (e.g. switch zoning), which is required.

The whole purpose of virtualization—higher efficiency, scalability and automation of operations, and cloud-like flexibility—is compromised when using FC-SAN for storage, because Fibre Channel is expensive, inflexible, and inefficient as storage for virtualization.

Containers, Dockers and Kubernetes

Containers effectively virtualize and run each application on top of the host operating system without duplicating the entire operating system stack for each instance. This makes containers more efficient than virtual machines, and today many large-scale cloud deployments have switched from virtualization to containerization.

Each container is a package of an application and its dependencies that run with a container management engine. This allows isolating applications and their dependencies from other applications running on the host machine while removing the need to run a separate operating system for each application, allowing for higher resource utilization and lower costs.

However, containers were initially difficult to deploy. Docker and then Kubernetes revolutionized container technology by providing a toolset to automate the creation and management of container images and their applications. Docker files contain an image of the application's source code and can quickly be built by specifying a list of commands. The docker builder packages the commands which, when ran, are used to build and launch the Docker container/application.

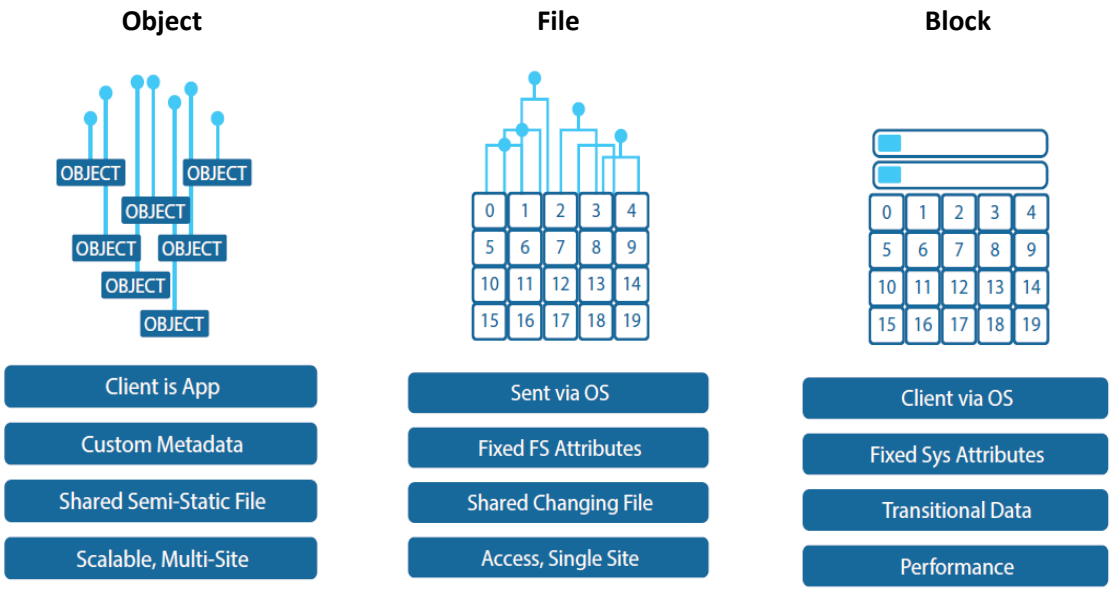
Likewise, Kubernetes is an open-source container management platform originally developed by Google and now managed by the Cloud Native Computing Foundation. With Kubernetes, a cluster of machines can be unified into a single pool of computing resources and applications are organized into groups of containers simplifying and automating the entire process of deployment and management.

Containers play an essential function in software-defined data centers due to their efficiency advantages over virtual machines, and for running microservices in the emergence of AI/DL/ML as new applications are built from the ground up.

Object Storage

Object storage is a data storage architecture that manages data as discrete entities called objects, as opposed to other storage architectures like file systems that manages data as a file hierarchy or block storage, which manage data as blocks. Object storage can only logically be accessed by Ethernet (or InfiniBand) and cannot be accessed by Fibre Channel.

In the past, object storage was usually deployed for massive volumes of data that did not require rapid access. But now, analytic tools such as Spark, Presto, and Hive—as well as the use of object storage for some types of AI, are driving the need for high-performance object storage. The performance and scalability of the networking fabric is critical in both use cases to deliver the scalability, high throughput, and low latency that may be required for different types of modern applications. Considering the volumes of data that must be ingested, stored, and analyzed, it quickly becomes evident that the storage architecture must be both highly performant and massively scalable, both traits of an ESF.



Data Analytics

The growth in Big Data, artificial intelligence (AI), deep learning (DL), and machine learning (ML) workloads has accelerated as organizations realize the economic value of data and analytics. Data analytics places a high demand on the underlying IT infrastructure. For example, data ingestion impacts capacity and throughput requirements, data streaming from databases require incredibly high I/O rates, and model training necessitates large volumes of rapid random reads of small blocks of data. To accomplish this, increased compute and GPU power are essential along with faster storage and a high-performance fabric with hardware-based accelerators and offloads including RDMA (RoCE), network virtualization, and data encryption. Without these, analytic data workloads suffer from reduced efficiency and performance.

Data analytics is also causing a shift to infrastructure that can dynamically connect underlying resources on the fly to match the requirements across the varying stages of the data pipeline: ingest, preparation, training, and inference. The ability to deploy underlying resources on the fly allows for matching the demands of each varying stage of the data pipeline. And, the ability to compose resources to match the demands of a variety of workloads allows for cost-effective scaling and increased resource utilization across the deployment of workloads.

The switch is at the heart of the network fabric and must be able to keep pace with fast and intensive data movements between compute and storage servers, including between CPUs, GPUs, memory, and storage. Mellanox Spectrum Ethernet switches provide 100/200 GbE line-rate performance and consistently low latency with zero packet loss. Spectrum is also the only RoCE-ready switch that can deploy RoCE effortlessly, offer end-to-end automatic RoCE acceleration and real-time RoCE visibility for easy troubleshooting.

Using NVMe-oF allows for accessing remote data at almost the same speeds as accessing local data. Mellanox ConnectX™ and BlueField DPU adapters support NVMe-oF, which utilizes RDMA for more efficient and lower latency data transfers. NVMe-oF is a high- performance storage protocol

specifically designed to take advantage of faster flash storage over RDMA transports such as RoCE, which allows remote direct memory access over Ethernet. Using Mellanox SmartNIC adapters and Spectrum switches to accelerate RoCE provide for a more efficient and faster way to move data between networked computers and storage while lowering latencies and CPU utilization.

Many of the hyperscalers have been reaping the benefits of data analytics for the last decade. The results they have shown through predictive analytics have optimized supply chain ordering, online advertising, and personalized recommendations for online shoppers. A move to an ESF enables sophisticated data analytic workloads by facilitating an infrastructure that is programmable, adaptable, and scalable. An ESF allows data analytic projects to scale efficiently by providing a high-performance, dynamic, and elastic infrastructure to handle the scale, performance, latency, and capacity demands of data analytics.

New Applications

Analysts suggest that significant spending within IT over the next few years will focus on migrations to cloud computing, investments in Big Data to glean more from their existing data sets, and the embedding of AI and ML methodologies into the very core of organizations. Applications that power these types of solutions are new to the IT industry. While they are fundamentally more flexible and agile than traditional enterprise applications, they have a material effect on the way data centers are designed, bringing with them new networking challenges, a need for higher bandwidth, and place unprecedented strain on the underlying network. Fortunately, Ethernet has kept pace with these new application requirements with increases in performance and capacity.

Cloud

The cloud application model relies on remote servers within the cloud for processing logic that is accessed through the web or internet connection. Thanks to cloud architectures, administrators can deploy applications more quickly and respond to problems without needing to take entire servers offline, and users can access and update files from anywhere. There are three main types of cloud infrastructures; Public, Private, and Hybrid. In a public cloud model, a service provider makes resources, such as virtual machines (VMs), containers, applications and/or storage, available to the general public. Private clouds are designed to offer the same features and benefits of public clouds but run behind a corporate firewall, usually for just one company's users, and offer more control over data. Private clouds also remove some security, confidentiality, and regulatory compliance worries. In a hybrid cloud, cloud applications work in concert with public cloud-based and local or on-premise components. Within the hybrid cloud model, there are more deployment options, greater security, and more flexibility because applications and data can move between public and private clouds.

If you are thinking of migrating your data from a traditional IT infrastructure to a private or hybrid cloud-based platform, you'll need to consider an infrastructure that can adequately scale, support automation, and provide adequate bandwidth. Cloud technologies increase CPU utilization due to the processing of overlay network encapsulation, Open vSwitch (OVS) packet processing, and high-performance virtualization and other intense workload processes. Similarly, cloud architectures increase east-west traffic on a network. This can waste expensive CPU cycles, clog network paths on switches that are not cloud-ready, and ultimately leaves a lot of resources underutilized--the result is that clouds, and their applications, become inefficient. Due to these challenges, data center

administrators must look for ways to implement intelligent, flexible networks that can provide enough bandwidth for application and storage requirements, all while alleviating CPU loads to enable cloud efficiencies and scale. A Mellanox-powered ESF offers end-to-end intelligent networking components that offload many of the networking tasks, thereby freeing CPU resources to serve more users and process more data and provide the bandwidth necessary to ensure users receive the best cloud experience.

Big Data

Organizations realize the economic value of data and analytics in gaining a competitive edge, which in turn is driving the growth of Big Data. Big Data is nothing more than a catchphrase used for a massive volume of both structured and unstructured data. This data is often so broad that it is difficult to process using traditional database and software techniques. Adequately named, Big Data's goal is to analyze extremely large data sets to reveal patterns or trends that an organization can use for growth or to maximize profits.

In most enterprise scenarios, the volume of data exceeds current processing capacities. The infrastructure required to run Big Data analytical applications (commonly Hadoop running on Hortonworks, MapR, or Cloudera, or in some cases on a NoSQL database) must be robust, scalable, and highly performant. The volume of data grows quickly, so the network infrastructure needs to scale accordingly to meet increasing requirements for throughput and storage capacity. Infrastructure matters; scalability, parallel processing, low-latency, and storage bandwidth are all necessary to properly leverage and optimize data analytics. Mellanox ESF I/O capabilities include low latency, high-throughput, low CPU overhead, and Remote Direct Memory Access (RDMA) to optimize Big Data application efficiency and scalability. The Mellanox ESF can accelerate Big Data analytics, allowing them to respond faster and enable higher scalability with linear performance gains. This equates to more job/sec and lowering total cost of ownership for Big Data applications. After all, the largest data sets and the finest software applications in the world won't run sufficiently without optimal infrastructure.

Artificial Intelligence and Machine Learning

Artificial intelligence and machine learning place a higher demand on the underlying IT infrastructure. For example, data ingestion impacts capacity and throughput requirements, data streaming from databases require incredibly high I/O, and model training necessitates rapid reads of large volumes of random small blocks of data. To accomplish this, increased compute and GPU power are essential along with a high-performance fabric with accelerators and offloads, and an ability to deploy underlying resources on the fly to allow for matching the demands of each varying stage of the data pipeline. Without these, AI and ML workloads suffer from reduced performance.

To scale AI and ML projects efficiently and provide sufficient performance, a dynamic and elastic infrastructure is required. An ESF can provide the scale, performance, latency, and capacity demands as well as deploying underlying resources on the fly to match requirements across the varying stages of the data pipeline; ingest, preparation, training, and inference. An ESF utilizes RoCE and NVMe-oF to offload the data transfer functions to the network adapter and bypass the CPU, and the switch keeps pace with the fast and intensive data movements between compute and storage servers. An ESF can

also match the demands of a variety of workloads, this allows for cost-effective scaling and increased resource utilization across the deployment of AI, and ML workloads.

The Evolution of Ethernet

Multiple data center trends are promoting the rapid evolution of Ethernet: the continued growth of virtualization and cloud computing which bring with them new networking challenges, the increasing speeds of server processors and the evolution of Flash and NVMe based storage arrays which continue to drive a need for higher bandwidth and new workloads that places unprecedented strain on the underlying network. Ethernet has kept pace with these data center trends while continuing to drive cost and operational efficiency and matching the increasing demands for performance and capacity.

While workloads are moving to the cloud and enterprises transform their on-premises IT infrastructure to emulate the cloud, they are realizing, as all major cloud providers long ago realized, that Fibre Channel is too expensive, too inflexible, and too limited as a storage network for their highly-scalable, super-efficient deployments. Hence, all the public clouds run both compute and storage on Ethernet (except for those that need the highest performance and efficiency and run on InfiniBand). As large and small enterprises deploy more virtualization, more containers, and more hyperconverged infrastructure to increase their flexibility and agility, they are following suit and deploying storage on Ethernet. Fibre Channel sales are essentially confined to the current installed base of larger enterprises like financial institutions, telecoms, cable operators, and government — those that were willing to pay for an expensive, dedicated, on-premises storage network.

Faster Speeds

Organizations are pushing more and more traffic through their enterprise data centers, and larger pipes are needed to accommodate the movement of all this data. To build out networks that can handle the increase in traffic and application workloads, the deployment of 10 GbE from the server has been increasing over the last few years. However, now 10 GbE is often not enough, requiring multiple 10 GbE ports per server to prevent bandwidth problems. This means organizations would need a significant increase in capital and operating expenses to install multiple new 10 GbE adapters and extra switch ports along with additional cables, rack space, power and cooling to support the required network bandwidth.

To better address high-performance networks and scalability requirements, a new 25 GbE standard was developed and was quickly adopted. The 25 GbE specification is more efficient than the 10 GbE specification. It delivers 2.5 times more bandwidth and utilizes a single-lane for 25 GbE, dual-lanes for 50 GbE, and quad-lanes for 100 GbE. In comparison, a single 40 GbE link requires four lanes so is less efficient, significantly reducing switch port density and increasing the cost of cabling and optics. 25 GbE allows for a smooth migration path to higher throughput using 100/200 GbE and increases scalability to accommodate growing workloads.

Sophisticated Traffic Management

Ethernet has earned pervasive adoption in the enterprise and cloud because of its speed, reliability, flexibility, and simplicity. The growth of IoT, social media, streaming video and other new technology innovations are flooding current network capacity and thereby creating a need for faster connection and increased capacity.

Why is Ethernet being relied on more than other networks to expand network capacity? First, the pervasiveness of Ethernet has proven a simple solution for connecting everything. Second, Ethernet is not a stagnant technology but a constantly evolving standard. With new traffic management mechanisms like traffic prioritization for QoS, and DCB capabilities --ETS, PFC, and QCN—as well as sophisticated congestion management options, Ethernet allows for more sophisticated traffic shaping than other network protocols. Traffic shaping is used to optimize performance, improve latency, and increase bandwidth for specific traffic flows. This is helpful for I/O-hungry or latency-sensitive applications as their traffic can be prioritized while non-essential traffic is sent on a best-effort basis.

Mellanox Spectrum-2 and Spectrum-3 switches take this a step further by offering Adaptive Flow Prioritization to enhance flow control, avoid congestion, and offer reliable data delivery. Adaptive Flow Control identifies small latency-sensitive traffic and prioritizes it over larger traffic flows, which are less susceptible to latency. The results are more efficient utilization of bandwidth and an increase in reliability of data transfers between compute, storage, internal and external data centers.

Affordability

The ubiquity of Ethernet and long runaway of success is mostly a result of its relatively low cost in terms of hardware and ease of deployment, and this is especially true when compared to networks such as Fibre Channel. Ethernet has always been designed to be easily upgradable and backward compatible, eliminating the need for the expense to rip and replace. Similarly, every enterprise has already standardized on Ethernet within their data centers due to its robustness and popularity, meaning all data centers already contain the wiring infrastructure needed to support it, and therefore, there is no additional investment in cabling. Accordingly, enterprises have invested in Ethernet networking expertise, which in turn has driven a large and growing pool of educated network engineers, something that is missing in the Fibre Channel market.

Ethernet has a large group of vendors all competing to develop and enhance products. This competition maintains a steady pace of new innovations and improved products, which constantly drives costs down, unlike the vendor duopoly, low volumes, and high price premiums found in the Fibre Channel market. Finally, the simplicity of Ethernet management provides for reduced overhead costs as well.

Offload Technologies and Network Accelerators

Modern data centers need to cope with the explosion of data that tends to suck up huge amounts of computing resources. Modern Ethernet implements several offload technologies that offload important networking, security, and storage tasks from the CPU and enable optimal data center efficiency. These are not features ingrained within the Ethernet transport, but technology embedded within Mellanox ConnectX intelligent NICs and BlueField DPUs. Mellanox SmartNICs offer advanced accelerations including packet queuing, stateless TCP offloads (like large receive offloads or LRO), RDMA over Converged Ethernet (RoCE), NVMe over Fabrics, VXLAN overlay, and virtual switch offloads and others. The physical network adapter offload capabilities can have a significant impact on network and storage performance. The difference is even greater under intensive workloads and in virtualized data centers where virtualization can introduce performance penalties by chewing up CPU resources. These acceleration technologies free up CPU resources for the compute tasks, allowing for improvements in performance, scalability, and efficiency within the data center. More information about a few of the offload and acceleration technologies follow.

Remote Direct Memory Access

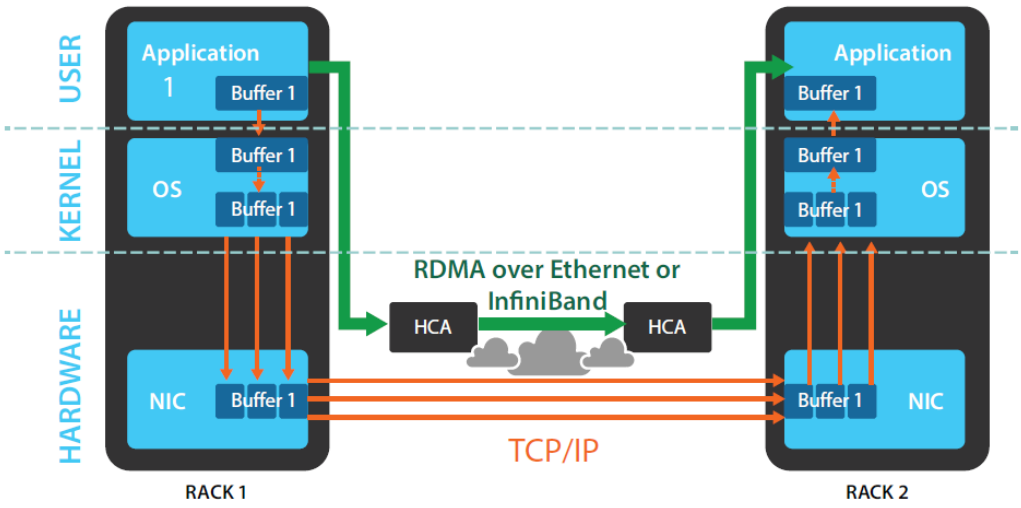
Remote Direct Memory Access (RDMA) is a technology that allows computers to exchange data memory-to-memory without involving the CPU, cache, or the operating system kernel. RDMA improves throughput performance by eliminating latency-costly data copying, processor interrupts, and context switching. This frees up valuable CPU resources so they can be used for application workloads. This also facilitates faster data transfer rates for both application and storage traffic.

Many of RDMA's practical uses include: storage applications where it is used to meet the increasing need for faster storage or to match the performance of newer SSDs; to accelerate server-to-server data movement, where it is useful in high-performance computing (HPC) environments; in AI and ML it is used to eliminating latency-expensive data copying, CPU/GPU interrupts, and context switching; and lastly it is used in Big Data workloads and today's complex virtualized cloud models to remove latencies. Essentially, RDMA is an excellent way to boost the performance of critical applications, and currently it's only available on Ethernet and InfiniBand networks; it's not supported on Fibre Channel.

RDMA Basics

RDMA has long been significant within the HPC community where it has proven itself, and its use is now growing within virtualization database, AI/ML, big data, and storage solutions. RDMA provides mechanisms to move data out of user space operations to a registered user space memory buffer without involving the kernel, network protocol stack or copying data from kernel space to user space. The data transfer mechanisms include Sends, RDMA Writes, RDMA Reads, fetch, add, compare and swap. And by offload or bypassing the kernel network stack, RDMA eliminates interrupts, and context switching, and utilizes kernel bypass for zero-copy operations. The following shows a graphic of data transfers with (orange) and without (green) the use of RDMA.

Figure 5. RDMA data moving from one application user space to the next without OS involvement



The benefits of introducing RoCE in data center infrastructure provides:

- Lower cost of ownership, because no separate storage networking infrastructure is needed
- Higher ROI across traditional and modern agile infrastructures
- Improved overall CPU utilization for running applications
- Efficient host memory usage
- Higher throughput and lower latency for compute and storage traffic

RoCE

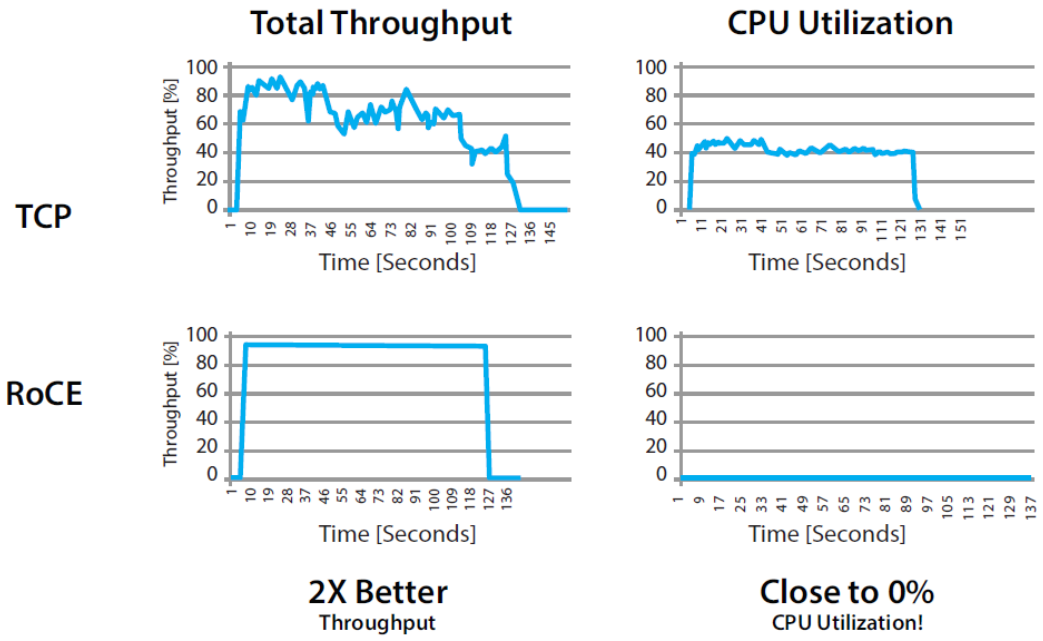
RDMA over Converged Ethernet (RoCE) is the most popular and most efficient way to run RDMA over Ethernet. It allows direct memory access over an Ethernet network and delivers superior performance because of lower latency and higher utilization of network bandwidth. Traditional SAN networks are unable to keep up with the demand of today's SSD flash-based storage and real-time applications. RoCE provides a positive impact by providing a faster network to complement today's faster storage. For example, Windows Server 2012 and later versions have standardized on the use of RoCE for accelerating and improving the performance of SMB file-sharing traffic and live migration. Other storage solutions examples iSCSI Extensions for RDMA (iSER), NVMe over Fabrics (NVMe-oF), and NFS over RDMA, all of which run over RoCE.

The offload of data transport tasks to a RoCE adapter frees up expensive CPU cycles to run application workloads. For many, these improvements in average latency and CPU utilization alone are enough to motivation them to move to RoCE-based networking.

Lately, RoCE is gaining momentum and being adopted by OEMs and many applications. A few of the applications benefiting from RoCE include Microsoft Storage Spaces Direct (S2D), VMware vSphere and vSAN, Spark, Hadoop, Oracle RAC, and many AI/ML frameworks. In this context, RoCE has currently become the eponym of low-latency networking in data centers and cloud deployments. As a bonus, the newest RoCE SmartNICs and DPUs also perform critical virtualization, cloud, security and

storage offloads, providing a double boost to performance and efficiency in and Ethernet Storage Fabric.

When comparing RoCE to iWARP, the other option for RDMA on Ethernet, it's clear that RoCE is the winner—it is faster, more scalable, supported by more network vendors, supported by more software applications, and the RDMA transport chosen by several hyperscaler companies (none of which have chosen iWarp). So, it is clear that RDMA provides big advantages for storage—and other types—of networking and that RoCE is the best way to run RDMA on Ethernet.



Zero Touch RoCE

Originally, RoCE required using switches that support Data Center Bridging (DCB), specifically using priority flow control (PFC) to deliver lossless Ethernet. Some customers were concerned about configuring DCB options on their switches, either because their network administrators were not familiar with it or because they felt changes like PFC were too disruptive. However, an ESF built with modern SmartNICs does not rely exclusively on this. For example, a Mellanox-powered ESF can leverage enhanced congestion mechanisms to eliminate the need for PFC and DCB. The newest Mellanox adapters also support the option to run RoCE with zero changes to the switch settings, a solution called Zero-Touch RoCE (ZTR). So now RoCE is available to accelerate storage and compute traffic on any kind of Ethernet network, without any special configuration required.

iSER

The iSCSI Extensions for RDMA (iSER) is a computer network protocol that extends the Internet Small Computer System Interface (iSCSI) protocol to use RDMA. The motivation for iSER is to use RDMA to avoid unnecessary data copying on targets and initiators. The traditional iSCSI protocol carries SCSI commands over a TCP/IP network. The iSCSI protocol traditionally encapsulates commands and assembles the data in packets to be sent over TCP/IP. iSER differs from traditional iSCSI as it replaces

the TCP/IP data transfer with the RDMA transport. By using the RDMA, the iSER protocol can transfer data directly between the memory buffers, eliminating network packet processing and data copying, which reduce latency and CPU load.

NVMe-oF Explained

Originally, SSDs relied on SATA or SAS for the bus interface and on Fibre Channel or iSCSI for networking block storage. Over time, SSD speeds have increased, but mechanical hard drives and the bus interface to them have stayed relatively stagnant. The older bus interfaces were developed to connect slower hard disk drives (HDDs), and as the performance of SSDs increased (DRAM transfer rates of 2500 MT/s), legacy bus interfaces proved to be inadequate to keep pace with the performance of SSDs. NVMe is the standardization of a new interface designed to bridge the gap between faster SSDs and the spinning media of HDDs, and NVMe over Fabrics (NVMe-oF) is the new way to network block storage based on NVMe SSDs. There are several adaptations to the NVMe specification to leverage ubiquitous protocols such as Ethernet and RoCE. The below section covers this in more depth.

Non-Volatile Memory Express

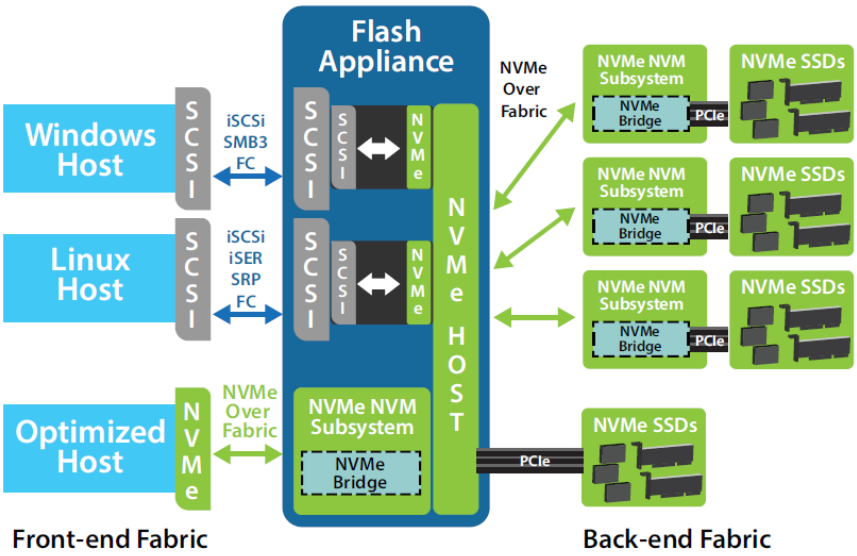
NVMe is the actual device interface specification for accessing non-volatile storage media—like SSDs--attached via a PCI Express (PCIe) bus. NVMe takes advantage of the low latency and fast access speeds of newer solid-state storage devices, offering a direct link to the CPU and networking devices via the PCIe bus, which reduces latency. The specification also requires fewer commands to complete I/O operations and supports multiple I/O queues, all of which can reduce CPU overhead and free up additional bandwidth.

NVMe-oF

NVMe over Fabrics is an extension of NVMe over the network and offers an efficient way of scaling devices over different fabrics, including InfiniBand, Ethernet, or Fibre Channel. This effectively extends the distance within a data center over which NVMe host devices and remote NVMe storage subsystems can be connected. NVMe-oF does this without adding much latency, and networked NVMe-oF devices have shown similar latencies as direct attach storage. NVMe-oF does require a fast, efficient network, and needs RDMA for the best performance.

For HPC, AI and machine learning customers, the fabric is usually InfiniBand, and for cloud customers it’s Ethernet. But for enterprise customers, it’s a choice between Ethernet (NVMe-oF on RoCE or NVMe-oF on TCP) and Fibre Channel (FC-NVMe).

Figure 6. Extend efficiency of NVMe over Front and Back-end Fabrics Enables efficient NVMe end-to-end model (Host <-> NVMe PC e SSD)



NVMe over RoCE

As mentioned earlier, RDMA permits the processing of network traffic to be handled by the NIC and bypass the software stack, resulting in higher throughput and better performance with lower latency. NVMe-over RoCE enables servers to access NVMe drives over standard Ethernet by creating a data fabric that uses RDMA.

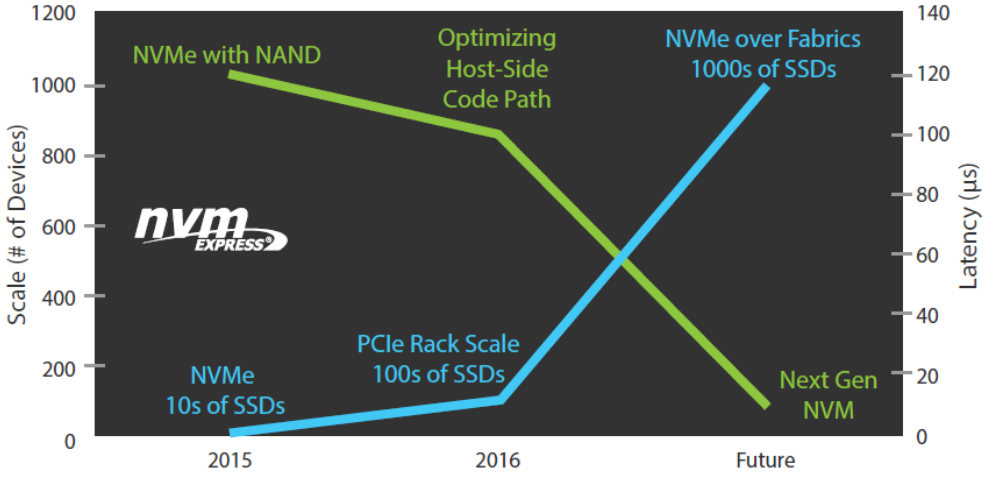
Since data centers mainly use Ethernet, we see RoCE as the underlying interconnect of choice. Recent debates have been sparked about which transport method, NVMe over RoCE using an Ethernet or FC-NVMe, is the best choice in the enterprise.

RoCE delivers many advantages over Fibre Channel when choosing a storage fabric:

- Higher performance of Ethernet (25, 50, 100 and 200 GbE) vs. Fibre Channel (mostly 16 G and a little 32 G, which actually run at 14 Gb/s and 28 Gb/s speeds)
- Lower cost—the per-port cost of 100 GbE is typically one-third that of 32 G Fibre Channel
- Ability to support all types of storage including block, file, object and HCI—whereas FC can only support block storage
- Flexibility of an ESF to run as a separate storage network or on one converged network that carries compute, storage and management traffic—FC can only carry storage traffic
- More advanced management, monitoring and security tools available for ESF, thanks to the much broader deployments and higher number of vendors for Ethernet

NVMe over TCP

Recently, NVMe-oF has added TCP as a transport option. NVMe over TCP transports NVMe packets inside TCP datagrams over Ethernet as the physical transport. Due to the use of Ethernet, NVMe over TCP can work on any network that can transmit TCP/IP traffic. Although both NVMe over RoCE and NVMe over TCP use Ethernet, NVMe over TCP encapsulates messages like FC-NVMe, as opposed to using frame-based transport like NVMe-oF on RoCE. As a result, NVMe over TCP is slower (higher latency) and requires more CPU resources from the servers and storage controllers than using RoCE. The fact that TCP is well understood and doesn't require any specialized fabric does make NVMe over TCP a tempting option for organization that are not ready to optimize their network infrastructure, and one of the big advantages of an ESF is that it can support NVMe over RoCE, TCP, or both. Some customers are already deploying both, using NVMe-oF on RoCE for servers and applications that require the fastest performance and NVMe over TCP (or iSCSI) for servers and applications that merely require good performance.



Ethernet Storage Fabric – A Deep Dive

An Ethernet Storage Fabric, or ESF in short, is the fastest and most efficient way to network storage. It leverages the speed, flexibility, and cost efficiencies of Ethernet. It exploits the best switching hardware and software, RDMA-enabled adapters with storage accelerators, and cables with the lowest bit error rate. It comes packaged in ideal form factors to provide performance, scalability, intelligence, high availability, and simplified management for storage. An ESF is optimized for scale-out storage environments and is designed to handle bursty storage traffic, route data with low latencies, provide predictable performance to maximize data delivery and support storage aware services. These are all crucial attributes for today's business-critical storage environments. An ESF can support new, faster speeds, including 25, 50, 100 and even 200 Gbps.

Additional ESF attributes include support for not just block and file storage, but also for object-based, along with storage connectivity for the newest NVMe over Fabric arrays. Additionally, an ESF must provide support for storage offloads, such as RDMA, to free CPU resources and increase performance. Not only is an ESF specifically optimized for storage, but it also provides better performance and value than traditional enterprise storage networks.

Predictable Performance

An ESF must keep up with the fastest storage technologies and a wide variety of business-critical applications. Many simply assume that the underlaying network always performs reliably and predictably. But it turns out that at the highest network speeds and with bursty traffic such as storage, predictable performance is extremely hard to deliver. Using a standard data center switch not designed for the demands of an ESF can result in higher and unpredictable network latencies.

The speed of today's NVMe SSDs, like 3D XPoint, have achieved latency of 10 microseconds or less. At this level, a few hundreds of nanoseconds of network latency significantly impact storage and application performance, especially if traversing multiple switches. Standard data center switches often produce latencies in the tens of microseconds. An ESF built with Mellanox Spectrum switches have ~300 ns port-to-port latency and zero packet loss, regardless of frame sizes and speeds. Furthermore, Spectrum switches are designed with a shared buffer, resulting in maximum micro-burst absorption capacity. Mellanox Spectrum switches are the only ESF switch that can support these fast storage types (refer to Tolly Report for more details).

ESF must be a transparent network fabric for scale-out storage, which means that access to remote data offers almost the same performance as access to local data, from the application's perspective. This translates into close-to-local predictable latency, line-rate throughput with QoS, and linear

scalability to accommodate dynamic, agile data movement between nodes – all simply and cost-effectively. An ESF constructed fabric built on Mellanox components meets all these requirements.

Simplified Leaf/Spine Architecture

To overcome architectural shortcomings, modern data centers are adopting the Leaf-Spine architecture for scale-out storage. The Leaf/Spine architecture has a simple topology wherein every leaf switch is directly connected to every spine switch, and any pair of leaf switches communicate with a single hop, ensuring consistent and predictable latency. By using Open Shortest Path First (OSPF) or Border Gateway Protocol (BGP) with Equal Cost Multi-Pathing (ECMP), a Mellanox ESF can utilize all available links and achieves maximal link capacity utilization. When network traffic increases, adding more links between each leaf and its spine has the capacity to provide additional bandwidth between leaf switches to avoid oversubscription, this helps avoid congestion and latency. Furthermore, as more and more scale-out storage deployments take the hybrid cloud approach, using Layer-3 protocols with standard-based VXLAN/ EVPN will seamlessly scale Layer-2 storage domains across data center/cloud boundaries with performance, mobility, and security, to ensure business continuity.

Scalability and Agility

Mellanox Spectrum Ethernet switches support all speeds, including 10/25/40/50/100 and 200 GbE. Meanwhile FC is just transitioning from 16 to 32 GFC. It's currently trails Ethernet speeds and will continue to lag with 400 GbE releasing in 2020 and 128 G Fibre Channel only on the drawing board for 2021 or 2022. The same Spectrum ESF switches that are deployed today will continue to service your needs when you migrate to next-generation storage or HCI platforms that require higher speeds. While with FC, speed increases will require existing switches to be replaced as new speeds are introduced.

Additional scalability advantages of an ESF come with the deployment of the suggested Leaf/Spine topology–TOR (leaf) and aggregation (spine). With the Leaf/Spine architecture, congestion, increased latency, and unpredictable performance caused by traffic jams in the traditional three-tier networks are eliminated. Within the data center, all storage I/O can transverse the ESF in a single hop (if the endpoints are in the same rack or three hops if across racks). This easily allows scaling from a half rack, to a full rack, to multiple racks, to multiple data centers. Dedicated ESF switches like Spectrum, are required to construct a storage fabric so that storage/HCI traffic, which includes bursty I/O's and data flows from faster devices such as NVMe SSDs, can always reach the destination within predictable response times. Using a switch not designed to meet the demands of storage can result in higher and unpredictable network latencies, even with a more efficient leaf-spine network designs.

As the leaf-spine architecture makes ESF extremely easy to scale, the ESF switches need to be simple to configure for fast and easy deployment and scale-out to accommodate agility. Automated network provisioning, monitoring, and management are required for virtualized workloads and storage traffic. So is seamless integration with clouds–including secure, isolated and agile work spaces for multiple tenants. Mellanox Spectrum switches allow for multi-tenant environments and bring cloud-scale performance and manageability to scale-out storage and HCI, including rich L2/L3 features, VXLAN/EVPN support for Data Center Interconnect (DCI), and OS-driven network automation.

Ideal for building storage and hyper-converged infrastructure clusters, Spectrum switches provide support for storage-specific monitoring, management, and automation through integration with storage arrays and storage management software. They also provide container support for storage-aware software. Better monitoring, better visualization, stronger GUIs, along with QoS and storage-aware features, make it better suited for storage than traditional data center switches. It provides better zoning and access control and can run containers, as well as other storage services making Spectrum the most agile storage switch available.

There are no limitations when using Ethernet, unlike FC, which has distance limitations due to its credit-based flow control. Having adequate buffer credits is essential to maintaining maximum I/O performance in a storage network, especially as distance increases. The problem being, FC switches have limited buffer credits. As an FC switch sends out credits, latency increases as the distance grows. This throttles the amount of data that can be sent as the sending switch has to wait for the acknowledgement of the received buffer credits. With Ethernet, TCP has a built-in error-correction facility, as TCP guarantees in-order delivery of packets and when using a Spectrum switch, packet loss is a thing of the past. Unlike most data center switches, which rely on retransmission when the network become congest and overload switches drop packets, Spectrum switches with their deep buffers and zero packet loss are far more reliable. Similarly, Ethernet has multiple features to assist with flow level congestion handling and QoS; ECN (Explicit Congestion Notification), PFC (Priority Flow Control), ETS (Enhanced Transmission Selection), and QCN (Quantized Congestion Notification). These features ensure flow control, avoid congestion, and offer reliable data delivery. Compared to FC with its credit-based link level control, which sustains throughput reduction when credits run low.

RoCE Ready

Mellanox, with its heritage in high-performance computing, leads in RDMA/RoCE technology development and offers the most mature and advanced RoCE solutions in the industry. By being the only vendor that offers a complete end-to-end RoCE solution, Mellanox enables RoCE at its best in any Ethernet network, regardless of speed, topology, and scale. The Mellanox ConnectX and BlueField SmartNICs provide zero-touch RoCE hardware-acceleration in the server, and Mellanox Spectrum and Spectrum-2 switches deliver RoCE optimization in the network fabric.

In storage environments, RoCE allows delivery of close-to-local latency for fast storage and in-memory applications such as NVMe-oF, Microsoft Storage Spaces Direct with SMB 3.0, Spark, and IBM Spectrum Scale (GPFS), to name a few. Many storage vendors including, Excelero, HPE, IBM, NetApp, Nutanix, and Pure Storage, all support RoCE for ESF environments. For example, Tencent achieved a record-setting performance using ConnectX RoCE capable adapters and Spectrum switches within big-data analytics workloads.

Table 1. 2016 Sort Benchmark: 2016 Sort Benchmark Contest Result

Sort Benchmark Competition	New 2016 World Records (Tencent Cloud)	Old 2015 World Records	2016 Improvement
Daytona GraySort	44.8 TB/min	15.9 TB/min	2.8X greater performance
Indy GraySort	60.7 TB/min	18.2 TB/min	3.3X greater performance
Daytona MinuteSort		7.7 TB/min	4.8X greater performance
Indy MinuteSort	55.3 TB/min	11 TB/min	5X greater performance

Not every network adapter or switch is ready for RoCE deployments as the use of RDMA requires hardware offloads on the NICs and intelligent traffic management on the switches. However, a Mellanox ESF contains all these and more. The optimized buffer design in Spectrum, combined with storage-aware QoS and faster Explicit Congestion Notification, help a Mellanox ESF delivers optimal congestion management for RoCE traffic. In the next few sections, we cover more on why a Mellanox ESF is uniquely positioned to help RoCE deployments accelerate storage traffic.

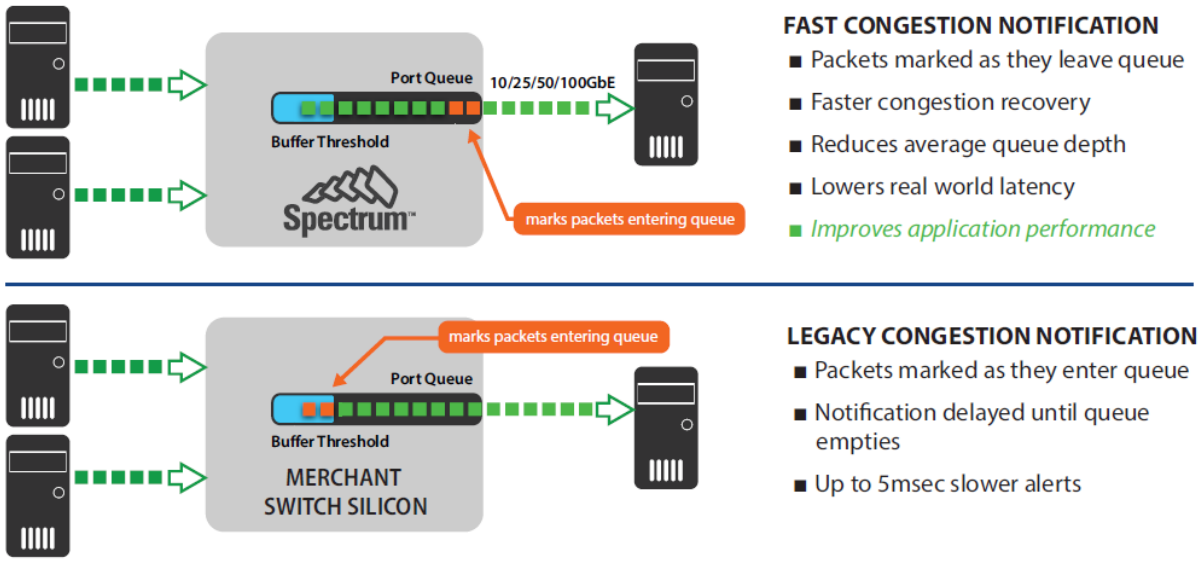
RoCE Acceleration

Storage-class persistent memory with microsecond latency SSDs on an NVMe shared fabric can shift bottlenecks to the network. The introduction of these fast storage technologies has changed the very concept of how a data center should be designed. More and more data center workloads are employing RoCE to ensure faster performance and delivery of more efficient CPU utilization.

Within a Mellanox ESF, RoCE acceleration first stems from the high-performance and low-latency Mellanox SmartNICs and their hardware-acceleration in the server. Next, acceleration from the Mellanox Spectrum and Spectrum-2 switches is delivered through RoCE optimization in the network fabric. Mellanox Spectrum switches provide line-rate throughput and ultra-low port-to-port switching latency at all speeds and packet sizes, with zero avoidable packet loss. Employing a shared-buffer architecture enables Mellanox Spectrum switches to deliver high performance and low latency fairly and predictably. This is crucial for software-defined platforms as running RoCE within defined priorities and policies can occur without disrupting the underlay switch characteristics. Both Mellanox SmartNIC adapters and Spectrum switches support RoCEv1 and RoCEv2, as well as Zero Touch RoCE (see Zero Touch RoCE section for more information).

Automatic RoCE acceleration also comes from innovations in advanced congestion control between the Mellanox adapter and switch. Mellanox Spectrum switches support per-flow congestion notification using Explicit Congestion Notification (ECN), Mellanox Spectrum switches offer a FAST ECN feature, which allows faster responses to congestion events. Once congestion is detected, instead of marking packets as they enter the queue (at the tail of the queue), Mellanox Spectrum switches mark packets when they leave the queue (at the head of the queue). As a result, the congestion notification is received up to milliseconds sooner. Earlier received alerts, in turn, reduces the chance of congestion occurring, and improves overall application performance. RoCE ready switches with optimized congestion management and QoS are required in an ESF switch to deliver a non-disruptive and transparent network fabric for business-critical applications.

Figure 7. Faster Congestion Notification by Mellanox RoCE



Effortless RoCE Configuration

Often, performance, configuration, and support issues in scale-out storage and HCI are network related. Depending on deployments, configuring the network fabric for RoCE may involve multiple steps – classifying ingress traffic flows, setting QoS for these flows, and enabling congestion control notification, for example (up to 25 steps with competitive products). Manually completing these steps for the switches is not trivial and often error-prone. In contrast, Mellanox Spectrum switches simplify the RoCE configuration with a single “roce” command, which applies a best-practices configuration for optimal performance. Mellanox offers a GUI for easy RoCE configuration with its network orchestrator, Mellanox NEOTM. With one click, NEO automatically configures RoCE on Spectrum Switches as well as ConnectX and BlueField SmartNICs for fabric-wide, end-to-end configuration, removing the possibilities of manual errors and support issues.

Zero-Touch RoCE provisioning is provided through Ansible integration and Mellanox’s network orchestration and management software, Mellanox NEO®. Ansible Playbooks and NEO not only improve operational efficiency but also eliminate network downtime caused by human errors. NEO provides network visibility, performance, and health monitoring, plus alerts/notifications to storage/HCI administrators, and guides them in troubleshooting. REST API-based, NEO can be easily integrated with scale-out storage or HCI software. For example, NEO is integrated with Nutanix AHV to provide automated VM-level network provisioning.

RoCE Visibility, Trouble Shooting, and Management

Real-time network telemetry is critical to manage, orchestrate, and troubleshooting/remediation of the network, especially for latency-sensitive RoCE data flows. With a single command, “show roce” Mellanox Spectrum switches provide advanced RoCE telemetry for real-time visibility of the RDMA traffic, including counters of RoCE traffic and non-RoCE traffic, congestion counters, as well as current and highwater buffer usage.

For further visibility, troubleshooting, and management, ESF solution leverage Mellanox What Just Happened. WJH is the Mellanox advanced streaming telemetry technology that provides real-time visibility into problems in the network. WJH goes beyond conventional telemetry solutions by providing actionable details on abnormal network behavior. Traditional solutions try to extrapolate the root cause of network issues by analyzing network counters and statistical packet sampling. WJH eliminates the guesswork from network troubleshooting.

The WJH solution leverages the unique hardware capabilities built into the Mellanox Spectrum and Spectrum-2 Ethernet switch ASICs to inspect packets at multi-terabit speeds – faster than software or firmware-based solutions. WJH can help in the diagnosing and repair of networks, including software problems. WJH inspects packets across all ports at line-rate, at speeds that overwhelm traditional Deep Packet Inspection (DPI) solutions. Analyzing each packet at line-rate, WJH can alert on performance problems due to packet drops, congestion events, routing loops, etc. If packets are dropping because of a bad cable or lousy optics, WJH will let you see those dropped packets and tell you why they were dropped. Congestion, buffer problems, or even security issues, WHJ can find them. For example, VM issues may be caused by ACLs dropping packets, WJH can identify a corrupted server or VM or poorly configured ACL.

In lossless environments, like NVMe over Fabrics (NVMe-oF) running on RoCE, you might have performance problems even though you are not dropping packets. The performance issues could be due to congestion issues or excessive pause frames or latency issues. Similarly, you might find out that the root cause is uneven load balancing across a LAG or ECMP group. Whether your problem is packet drops or poor performance without packet drops, WJH was built to get to the bottom and provide you the best streaming telemetry for superior network visibility. WJH is an Open Ethernet solution that can be integrated into open source tools like Grafana, and Kibana, but also works with turn-key data center wide monitoring solutions like Mellanox NEO.

Mellanox Onyx’s built-in automation infrastructure reduces operational expenses and time to service by minimizing manual operations and eliminating configuration and provisioning errors. Automation tools such as Ansible, SaltStack, ZTP, and Puppet enable you to automate fabric configuration and large-scale deployments.

Components of an Ethernet Storage Fabric

Low latency SSDs, persistent memory, and NVDIMMs are pushing the envelope for storage access times to under 1 microsecond. Thus, the network interconnecting compute to storage is more important than ever. With Ethernet at 100 Gb/sec today and on track for 200 and 400 Gb/sec in 2020 and Fibre Channel only at 32 G (28 Gbps) today, we see the speed gap between Ethernet and Fibre Channel continuing to widen. Ethernet is where the innovation is today, and Ethernet Storage Fabrics will soon displace Fibre Channel for the number one storage interconnect.

What kind of Ethernet products are required for an ESF? Standard-off-the-shelf NICs? Ordinary data center switches? No, and this is important because standard NICs don't have the offloads and accelerators required to ensure the CPU is running efficiently. And most data center switches drop packets, lose data or leave a gap in the transmission of storage data packets, which can easily overload a storage network as it tries to play catchup. Optimizing the network for high-bandwidth, low latency, and efficiency becomes a critical architectural element of the Ethernet Storage Fabric innovation. The below products make up an Ethernet Storage Fabric and more information about why they are the best to build out an ESF follows:

- Mellanox ConnectX® SmartNICs Ethernet adapters are RDMA-optimized, support Zero Touch RoCE plus storage, cloud, and security offloads (web link)
- Mellanox BlueField® SmartNIC Data Processing Unit (DPU) combines ConnectX adapters with advanced software and FPGA programmability (web link)
- Mellanox NVMe SNAP™ brings virtualized storage to bare-metal clouds and makes composable storage simple (web link)
- Mellanox Spectrum® Ethernet switches offer predictability, low-latency and high-bandwidth, are RoCE ready and provide zero packet loss for the best storage fabric (web link)
- Mellanox LinkX® cables provide an industry-low Bit Error Rate (BER), lower latency, and lower power for superior connectivity (web link)
- Mellanox Onyx® advanced Ethernet switch operating system optimizes ESFs by providing key automation, visibility, and management features that simplify storage fabric management (web link)
- Mellanox What Just Happened® (WJH) advanced telemetry provides real time visibility into network problems for root cause analysis (web link)

Mellanox ConnectX SmartNICs

The Mellanox ConnectX SmartNIC is a prime example of an adapter that is created for ESF environments. The ConnectX supports unique storage-centric features and offers hardware accelerators and offloads to ensure efficient CPU utilization on the host. By utilizing in-hardware acceleration for Remote Direct Memory Access (RDMA/RoCE and NVMe-oF) traffic bypasses the network stack, freeing the CPU for better workload efficiency. ConnectX adapters utilize the IBTA standards-compliant RDMA technology to deliver low-latency and high performance over Ethernet networks. This is accomplished without leveraging any Data Center Bridging (DCB) capabilities to provide efficient low latency data transfers over Layer 3 Ethernet. Eliminating the configuration process allows for easy installation and bring up of ConnectX enabled ESFs. Similarly, the use of NVMe over fabrics provides for the lowest latency storage connectivity possible, permitting network storage to demonstrate direct attach performance.

ConnectX adapters enable superior performance when using overlay networks by introducing hardware offload engines to encapsulate NVGRE and VXLAN traffic. Similarly, an embedded eSwitch and support for Open vSwitch virtual switching can extend ESFs throughout cloud, software-defined, and virtualized environments. ConnectX SR-IOV technology provides dedicated adapter resources by virtualizing I/O and guarantying isolation and protection within HCI and virtualized deployments. I/O virtualization provides data center managers better server utilization while reducing cost, power, and cabling complexity.

Applications utilizing ordinary TCP/UDP/IP transport can achieve industry-leading throughput over 25/50/100 and 200 GbE networks, further increasing bandwidth for latency-sensitive applications. All ConnectX adapters support Windows, Linux distributions, VMware, and FreeBSD, along with top hypervisors, including Windows Hyper-V, Xen, KVM, and VMware are compatible with configuration and management tools from OEMs and operating system vendors. Using Mellanox SmartNICs for ESFs permits for significant improvements over commodity NICs. The results are lower latencies, reduced CPU involvement, and higher data center efficiency resulting in lower operational costs.

Mellanox BlueField DPUs

Mellanox's family of programmable SmartNICs encompasses the advanced capabilities of the ConnectX network adapters with advanced software and FPGA programmability, providing data centers with levels of performance and functionality previously unseen in the market. These cards are based on the BlueField Data Processing Unit (DPU) – a family of innovative networking and I/O acceleration ICs. The new generation of SmartNICs is the perfect blend of hardware and programmable accelerations enabling secure and high-performance network solutions for cloud, storage, machine learning and edge computing applications while increasing productivity and reducing total cost of ownership.

BlueField high-performance, programmable networking engine enables the customization and optimization of both control and data path operations. Key features include an embedded virtual switch with programmable ACLs, transport offloads, stateless encapsulation and decapsulation of overlay protocols (NVGRE, VXLAN, MPLS), dedicated hardware offloads (including NVMe-oF), as well as superior RDMA and GPUDirect® accelerators.

Mellanox NVMe SNAP

Mellanox NVMe SNAP technology, delivered as part of the BlueField DPU family, allows customers to compose remote server-attached NVMe Flash storage and access it as if it were local. NVMe SNAP makes use of these existing NVMe interfaces, to give customers the composability and flexibility of networked flash storage, combined with the advantages of local SSD performance, management, and software transparency. This NVMe SNAP technology is combined with BlueField's powerful multicore Arm processors and virtual switch and RDMA offload engines, to enable a broad range of accelerated storage, software defined networking, and application solutions. The Arm processors in combination with SNAP can be used to accelerate distributed file systems, compression, de-duplication, big data, artificial intelligence, load balancing, security and many other applications.

Mellanox Spectrum and Spectrum-2 Switches

What kind of Ethernet switch is ideal for storage? First, it must be FAST, meaning high-bandwidth, non-blocking, and with consistently low latency. After all, faster storage needs faster networks, primarily to accommodate for all-flash arrays. The Mellanox SN-series of Spectrum and Spectrum-2-based Ethernet switches are designed and optimized for high performance, flexibility, and value within all storage environments. The SN-series has storage aware features and is RoCE ready with support for Data Center Bridging (DCB), including DCBx, Enhanced Transmission Specification (ETS), Priority Flow Control (PFC) and iSCSI traffic can be classified and prioritized using iSCSI-TLV. However, when used with Mellanox ConnectX or BlueField SmartNICs, enhanced congestion mechanisms are deployed to make RoCE lossless and fully compliant with IBTA industry standard RoCE without the configuration of DCB.

SN-series switches use an intelligent buffer design to ensure buffer space is allocated to the ports that need it the most. This ensures data I/O is treated fairly across all switch port combinations and packet sizes. Other switch designs typically segregate their buffer space into port groups which makes them up to 4-times more likely to overflow and lose packets during a traffic microburst. This buffer segregation can also lead to unfair performance where different ports exhibit wildly different performance under load despite being rated for the same speed. Latency is among the lowest (Spectrum 300 ns port-to-port and Spectrum-2 425 ns port-to-port) of any generally available Ethernet switch. The silicon and software are designed to keep latency consistently low and offer zero packet loss across any mix of port speeds, port combinations, and packet sizes.

The SN-series switches allow for maximum flexibility with unique port configurations allowing for two units to be deployed side-by-side for fault-tolerant solutions, perfect for small port count HCI deployments. With speeds ranging from 1 GbE to 200 GbE, they are ideal for flash storage environments and support all storage types; block, file, and object, as well as hyperconverged infrastructure (HCI). By maintaining high-bandwidth at low latencies and line-rate throughput across all packet sizes and ports, Spectrum and Spectrum-2 switch ensure predictable data delivery. Explicitly designed for storage workloads and ideal for building a scalable ESFs, they integrate with storage and network management tools and have the capability to run a container on the switch to provide storage-specific services.

Mellanox LinkX Cables and Transceivers

Massive demand for high-capacity data center infrastructures and high-speed interconnects are playing an increasingly important role in technology these days. This is placing more focus on the cables and transceivers that connect everything. Designing cables that work at blazing fast speeds of 25 G and greater and that operate for many years under high temperatures is not a trivial task. All Mellanox cables are designed to HPC supercomputing Bit Error Ratio (BER) standards, which requires a BER rating of one-bit error in 10^15th bits transmitted (expressed as 1E-15). The IEEE Ethernet industry standard is BER of 1E-12, which is one-bit error in 10^12th bits transmitted or about 1,000 more errors than Mellanox standard cables.

Mellanox offers one of the industry's broadest portfolios of 10, 25, 40, 50,100, 200, and 400 Gb/s Direct Attach Copper cables (DACs), Active Optical Cables (AOCs) and Transceivers, and 2-to-1 or 4-to1 break-out cables with reaches from 0.5 m to 10 km. Mellanox designs and manufactures all of its switch systems, network adapters, DAC and AOC cables, and optical transceivers. Not only the units themselves but also the integrated circuits (ICs) that go into the switches, network adapters, and transceivers –including advanced technologies such as Silicon Photonics technology. This vertical "end-to-end" integration approach ensures everything is designed to work optimally together. The vertical integration and manufacturing enable Mellanox to offer an unparalleled quality experience and ensures everything is interoperable. Why use Mellanox LinkX cables? More cabling options, exceptionally low bit error and low-latency ratings, real system tests, designed and manufactured by Mellanox.

Mellanox Onyx OS

Mellanox Onyx is a high-performance switch operating system, designed for the scale and demands of next-generation data centers. With built-in workflow automation, monitoring & visibility tools enhanced high availability mechanisms, and more, Mellanox Onyx simplifies network processes, increasing efficiency and reducing operating expenses and time-to-service.

Mellanox Onyx offers multiple best-in-class buffer utilization monitoring and enhanced quality of service (QoS) mechanisms. It also embraces advanced congestion avoidance and congestion management features, which are vital to unlocking the scalability and performance of applications from the worlds of storage and Artificial Intelligence. Whether building a robust Ethernet Storage Fabric (ESF), public or private cloud, customers can leverage the flexibility of Onyx to tailor their network platform to their environment.

What Just Happened (WJH)

For visibility, troubleshooting, and management, ESF solution leverage Mellanox What Just Happened (WJH). What Just Happened (WJH) is the Mellanox advanced streaming telemetry technology that provides real-time visibility into problems in the network. WJH goes beyond conventional telemetry solutions by providing actionable details on abnormal network behavior. Traditional solutions try to extrapolate the root causes of network issues by analyzing network counters and statistical packet sampling. WJH eliminates the guesswork from network troubleshooting.

The WJH solution leverages the unique hardware capabilities built into the Mellanox Spectrum and Spectrum-2 Ethernet switch ASICs to inspect packets at multi-terabit speeds – faster than software or firmware-based solutions. WJH can help in the diagnosing and repair of networks, including software problems. WJH inspects packets across all ports at line-rate, at speeds that overwhelm traditional Deep Packet Inspection (DPI) solutions. Analyzing each packet at line rate, WJH can alert on performance problems due to packet drops, congestion events, routing loops, etc. If packets are dropping because of a bad cable or lousy optics, WJH will let you see those dropped packets and tell you why they were dropped. Congestion, buffer problems, or even security issues, WHJ can find them. For example, VM issues may be caused by ACLs dropping packets, WJH can identify a corrupted server or VM or poorly configured ACL.

In lossless environments, like NVMe over Fabrics (NVMe-oF) running on RoCE, you might have performance problems even though you are not dropping packets. The performance issues could be due to congestion issues or excessive pause frames or latency issues. Similarly, you might find out that the root cause is uneven load balancing across a LAG or ECMP group. Whether your problem is packet drops or poor performance without packet drops, WJH was built to get to the bottom and provide you the best streaming telemetry for superior network visibility. WJH is an Open Ethernet solution that can be integrated into open source tools like Grafana, and Kibana, but also works with turn-key data center-wide monitoring solutions like Mellanox NEO.

Benefits of an Ethernet Storage Fabric

The real benefit of an ESF is how integral all components work together, specifically designed for storage workloads and perfect for building a modern fabric for storage workloads. An ESF is ideal for building fast and scalable storage networks for all storage types; block, object, and file storage and for structured and unstructured data. A wide range of faster storage, improved protocols, and storage offloads and accelerators are causing Ethernet to grow in popularity compared to other storage networks. Below are a few more reasons why an ESF is gaining momentum as the best storage interconnect:

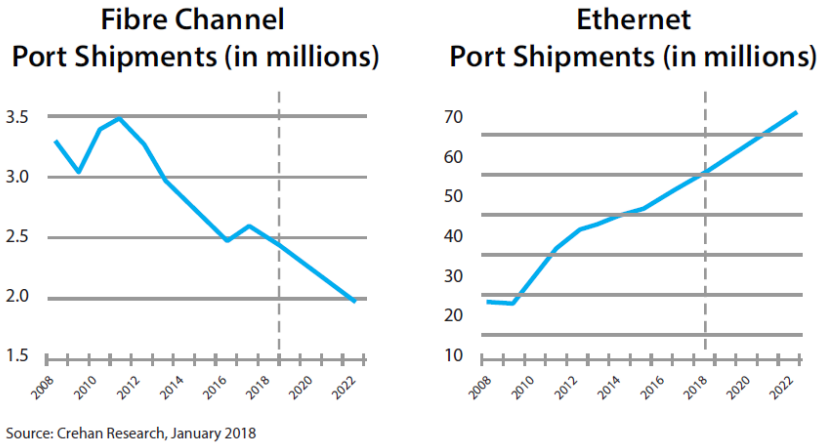
Fastest Performance: The ESF is a dedicated network fabric for scale-out storage and HCI. The congestion, increased latency, and unpredictable performance caused by traffic aggregation in the traditional three-tier network are now gone. Within the data center, any storage/HCI I/O transverses the ESF in a single hop if the endpoints are in the same rack, or in three hops if across racks. If dedicated ESF switches are used to construct the fabric (we will come back to this point later), storage and HCI traffic, including bursty I/Os, always reaches its destination with predictable response time. With RDMA over Converged Ethernet (RoCE) offload and native NVMe over Fabrics (NVMe-oF) acceleration, applications are serviced at the highest performance level, in accordance with SLAs or predefined policies.

Simple to Deploy, Manage, and Scale: Ethernet is ubiquitously used in data centers, and easy and rapid to expand. By converging all network and storage traffic within scale-out storage and HCI environments onto Ethernet, ESF eliminates network silos (such as Fibre Channel used with legacy SAN), resulting in a single network fabric to manage. Beyond the boundary of a single data center, the use of overlay technologies such as VXLAN/EVPN, which create efficiencies allowing expansion across multiple data centers.

Automation, Security, and Storage-aware QoS: An ESF provides automated network provisioning, monitoring, and management for virtualized workloads and storage traffic. Seamlessly integrated with clouds, ESF supports secure and isolated workspace for multiple tenants on scale-out storage and HCI. Combined with the intelligence in auto-discovering storage devices on the fabric and allocating proper network resources for storage-aware QoS, the ESF delivers a non-disruptive and transparent network fabric for business continuity of business applications.

Cost-Effective: Ethernet is a de-facto network in data centers and clouds. Extensive usage and high-volume shipments have driven down the hardware cost while ensuring rapid technology innovation and enterprise-class quality. Furthermore, innovative and scalable management tools and automation software for configuration, monitoring, and troubleshooting have grown out of the vast amount of Ethernet networks deployed by both enterprise and cloud customers. These management tools significantly reduce operational costs for managing scale-out storage and HCI. Easy application migration over a single fabric with automation tools maximizes uptime and resource utilization, also lowering operations costs.

Containers, Docker, and Kubernetes: The move to modern data centers is driving new and dynamic operation models. An ESF must also provide a wide range of tools to address these needs. For example, support for Docker containers, enabling software to be run in isolation. This provides fast and secure delivery of customized applications, giving customers a unique edge to quickly integrate and improve development cycles and share storage resources between containers.



Source: Crehan Research, January 2018

Comparing Alternatives

It turns out, Ethernet-connected storage is growing much more rapidly than FC connected storage, and about 80 percent of storage capacity today is well suited for Ethernet (or can only run on Ethernet). Only 20 percent of storage capacity is the Tier-1 block storage that traditionally goes on FC SANs, and most of that block storage is suitable to run on iSCSI or newer block protocols such as iSER (iSCSI RDMA) and NVMe over Fabrics (NVMe-oF, over Ethernet RDMA).

The demise of FC, as shown in the charts below, has continued its downward spiral, while Ethernet shipments have continued to rise. Technology developments improving storage over Ethernet has already taken place, including the arrival of mainstream RDMA over 25/100 GbE networks, the emergence of fast storage, such as NVMe and 3D XPoint SSDs, the growing adoption of object

storage, and the convergence to public, private, and hybrid clouds which all contributing to the success of Ethernet. Fibre Channel innovation has stagnated, and it remains a block-only storage solution deployed only in the enterprise; it has no use in the cloud, big data, or machine learning, and is cumbersome in HCI deployments due to the addition of a separate, disparate network.

Ethernet is faster than Fibre Channel in bandwidth and raw throughput, running at speeds of 25, 50, 100, or even 200 Gb/s, while Fibre Channel is lagging at 32 GFC rate. With the introduction of lossless transmissions, and support for RDMA connections to lower latency and frees up CPU cycles brings massive performance improvements to Ethernet shared storage. We also see the next advancements in storage networking starting to gain traction with the introduction of NVMe (both NVMe-oF and NVMe-TCP protocols), which could mean the end for Fibre Channel storage.

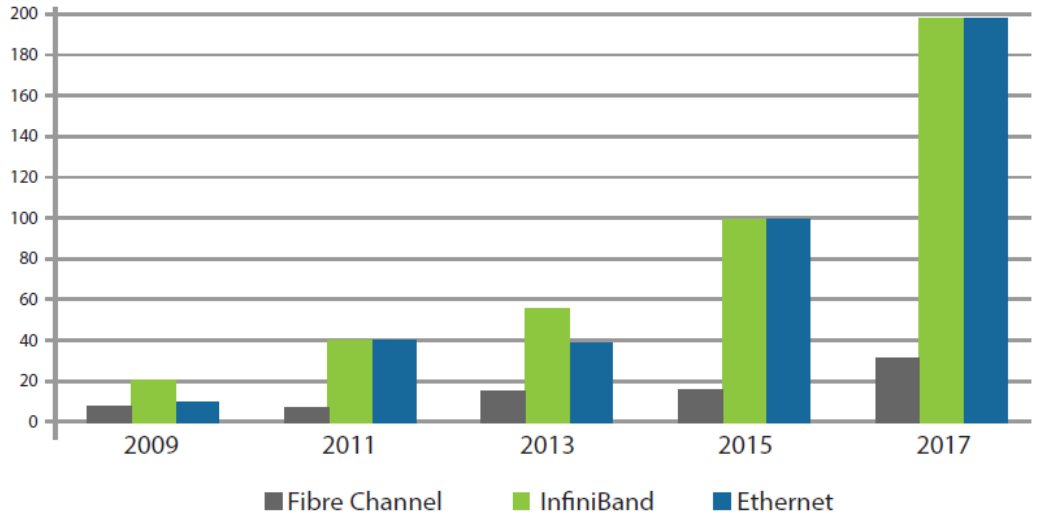
Ethernet easily supports multiple storage protocols—block, file, object, etc.—simultaneously, and allows client, server, and storage traffic to share the same network, using traffic prioritization and QoS. Even with 32 Gb/s FC, there will be significant limitations such as only support for block storage traffic.

Finally, Data Analytics, which encompasses Big Data, AI, ML and large clusters of compute that processes and analyze large sets of data, is a growing trend being driven by almost every industries including media and entertainment, oil and gas exploration, semiconductor design, automotive simulations, pharmaceutical research, finance, and retail. With data file sets that can be multiple terabytes and some of the largest exceed a petabyte. The perform data analytics on these large data sets requires clusters that are hundreds or thousands of nodes, all processing millions of operations per second. Storage must leverage clustered file systems, clustered file systems (Gluster, Lustre, Spectrum Scale), or object storage like Ceph, which mostly don't support Fibre Channel. Fibre Channel is too slow, too expensive, and too inflexible to play in any of these growing markets. The only growth storage category that's using Fibre Channel are all-flash arrays that are replacing older Fibre Channel storage. However, all-flash arrays are being connected more and more often with Ethernet storage protocols like iSCSI, iSER, and SMB Direct. In the meantime, FC currently only plays in a shrinking market far away from the primary storage growth.

Speed Discrepancies of Fibre Channel

In Ethernet, the network speed is appropriately named “25 GbE,” which refers to the throughput which is 25 Gb per second. Whereas in FC, the name “32 GFC,” which one would assume to mean the throughput is 32 Gb/s. However, the speed is significantly less than the names for each generation.

Figure 8. Deployable Network Speed in Gb/s



When FC was first introduced, it was named after its baud rate, followed by a switch from Gb/s to GFC and again in 2013 to its current naming association which has been switched to a generation-based naming format (Gen 5 = 16 GFC, Gen 6 = 32 GFC). Unfortunately, this nomenclature is very misleading as most individuals presume, Gen6 FC or 32 GFC is, in fact, 32 Gbps. But at 32 GFC, the gap is almost 5 Gbps, and in future generations, that naming gap will only continue to increase. So, when comparing 25 GbE to 32 GFC, one would assume FC is 7 Gbps faster than Ethernet, but a line rate comparison would prove that 32 GFC has an effective bit rate of only 27.2 Gb/s, and only slightly faster than 25 GbE.

The fact that the entire SAN industry refers to FC in terms of GFC or Gbps is misleading. Similarly, when we hear 128 FC is available and is the fastest storage interconnect. The FC pundits are misleading you again. First, 128 is only available as an interconnect for the largest director-class switches. Second, there is no 128 GFC support for storage arrays, let alone any software (vSAN, MSFT S2D, Veritas). For further information on the topic refer to this LinkedIn article.

Not all Fabrics are Created Equal

Ethernet has adapted to the recent technological advances within modern data centers and is well equipped to handle future data and storage networking trends. Ethernet is seen as a key technology enabler for companies to achieve their strategic goals. Ethernet has risen to the challenge by simultaneously being the ubiquitous protocol that can run the business, connect the world, and can be leveraged for innovative solutions to help drive growth. Ethernet allows for network and storage acceleration, the consolidation of compute and storage, and can achieve significant cost-performance advantages over Fibre Channel and multi-fabric (Fibre Channel/ Ethernet) networks.

While Fibre channel is only a one-trick pony, offering standard block access. Ethernet can be leveraged for Block, file, and object storage access. Fibre Channel is stuck networking block storage from the past, all while Ethernet continues to drive everything from cloud deployment models, data analytics and is posed to be the foundation for the Internet-of-Things and autonomous cars in the future. High speed, low latency, processor efficiency, and network accelerators are becoming a real differentiator for companies. Software-defined networking and network function virtualization will be used to build controllable, secure, distributed networks.

The growth of the Internet of Things (IoT), 5G wireless, and AI are all driving the move of compute to the edge. IoT means more – and smarter – devices are generating and consuming more data, but in the field, far from traditional data centers. Ethernet allows for SmartNICs, such as the Mellanox ConnectX family, to offload work from the CPU. With hardware-accelerated functions, they can handle overlay networks, RDMA, container networking, virtual switching, storage networking, and video streaming. They can accelerate the adapter side of network congestion management and QoS and can perform in-line encryption and decryption in hardware at high speeds, supporting IPsec and TLS. With these essential but repetitive network tasks safely accelerated by the NIC, the CPUs and GPUs at the edge connect quickly and efficiently, all the while focusing their core cycles on what they do best—running applications and parallelized processing of complex data. The future of fabrics is much different the specialized function of Fibre Channel adapters and a role that only Ethernet SmartNICs can fill.

Ethernet - 4X the Performance, ¼ the Price of Fibre Channel

Fibre Channel has long been thought of as the performance leader when it comes to storage interconnects. However, when Ethernet emerged with new storage aware features and released 25, 50, and 100 Gbps speeds, they closed the performance gap and directly challenged Fibre Channel's performance, and reliability as a storage interconnect. The next generation of Ethernet speeds releasing in 2020 (200 and 400 Gbps) will introduce significant performance improvements that will leave FC generations behind when it comes to speeds. On the host side, FC is currently moving to 32 GFC, which, if you recall from the Speed Discrepancies of Fibre Channel section above, it's roughly 27 Gbps speeds, while 100 Gbps Ethernet speeds are available for host today. That's approximately a 4X the performance of FC.

But what about costs? Fibre Channel is a pricey storage option which has been able to get away with keeping cost high due to the duopoly of two vendors who have no reason to discount products for competitive purposes. For this reason, Ethernet also holds a dramatically lower cost advantage.

A dual-port ConnectX-4 EN intelligent Ethernet adapter which supports speeds to 100 Gb/s resells for approximately $750, while a Broadcom dual-port LPE3200 32 Gb Fibre Channel adapter retains for $1499. Almost twice the price, but if you take into consideration the price/performance ratio, which refers to a product's performance divided by its cost, Fibre Channel is (1450/27) = $55.52 per Gbps while Ethernet is (750/100) = $7.50 per Gbps. A substantially lower price/performance ratio, making it much more desirable.

Lastly, if we look at roadmap and R&D spend, Fibre Channel has minimal development efforts other than speed transitions and accommodations for new interface types to support those speed transitions. Meanwhile, Ethernet has continually focused on improving features sets to address the latest industry trends. Ethernet has proved to be disruptive by successfully targeting new initiatives, gaining a strong foothold within them, and delivering more-suitable functionality—frequently at a lower price. As we see in the storage industry, Fibre Channel, while it is the incumbent, it's chasing higher profitability and tends not to respond vigorously to changes in the market.

Implementing an Ethernet Storage Fabric

The permutations of ESF designs are virtually infinite and Mellanox solutions are flexible enough to enable a wide variety of ESF designs and to support a wide range of ESF solutions. The key to a successful ESF design is to define the requirements thoroughly and to understand the ESF design variables. Understanding these will help in the selection of an appropriate design.

Usually, an ESF is built to solve a business problem. For example, the problem statement could be: "We need to optimize our storage utilization." Or, "our nightly backups don't complete within the allowable window, so we need them to run faster." When evaluating the following list of design considerations, keep in mind that the overriding concept is always the same: the ESF solution must solve the business problem that drove the aspiration for an ESF in the first place. In order to be effective, an ESF solution should:

- Solve the underlying business problem
- Provide an appropriate level of data protection consistent with the requirements of the business problem
- Provide the proper level of performance
- Meet requirements for availability and reliability
- Be able to be managed effectively
- Be scalable and adaptable to meet current and future requirements
- Be cost-effective

These and other requirements must be well thought out before the design is created. For inspiration on design options and details on ESF solutions, refer to the ESF Storage Solutions section below.

Moving from a FC SAN to an ESF

When moving from a SAN to an ESF, it could be part of a regular storage refresh cycle, or a leap into hyperconverged infrastructure. Data migration should be executed with a well-planned process so that it does not impact application availability. Within the modern data center there are no longer windows for scheduled backups or maintenance windows for downtime. Applications run 24×7 and can't be brought down for any length of time to complete a data transfer. There are also huge data sets that, even with advances in networking, are often too large to transfer overnight.

Requirements for Migration

In the modern data center, migrations need to happen while applications are still available, and it should also allow for a rapid switch- back if something goes wrong with the migration. Migration utilities have evolved to meet these requirements. New methods include specific copy utilities built into the operating system, generic replication software, or storage systems specific migration utilities.

Migration Options

Migration tools, replication software, and storage virtualization software are all options for migrating data from only SAN storage to new Ethernet storage arrays. Operating system utilities have the most significant advantage because they are often free. However, operating system utilities typically cannot perform these tasks in a non-disruptive fashion. While some can do file-by-file sync, they usually require application downtime to ensure a clean copy of the data gets transmitted. These utilities also consume most of the CPU and memory of the host when doing the copy.

Most storage vendors offer storage migration software, which is naturally the best route as data copy is completed east and west across the SAN to the ESF. The IP network of the ESF has plenty of bandwidth, and the data moves from one storage system to another, consuming minimal resources. The biggest issue for array replication software is that it's single purpose so it comes with a high acquisition cost.

Replication software can be a cleaner process, typical involving copying data at a block level to the connected host and then via an IP network to another host that is connected to the new ESF. Lastly, the second host then must copy the data to its assigned storage volume. Replication can be done in real-time and with minimal downtime. In this scenario, the replication software consumes CPUs and memory resources of both hosts and is burdened by having to copy data across from the SAN. It's also sending it across the IP network utilizing IP software stack, which is CPU intensive and not well suited for the migration of a large amount of data.

The benefit of replication software is that once completed, a synchronization of changed blocks in near real-time, which does not consume much overhead, can be completed asynchronously over the IP network. Replication also permits for failback options in the case of a failure.

Finally, storage virtualization software might be the ideal solution. It provides the foundational migration strategy that the data center needs while at the same time, allowing for operational and cost saving benefits of the virtualization of storage assets. A storage virtualization solution can copy data from the original SAN to the new ESF storage system, in an east-west manner, without involving anything from the servers. This can be done in two ways, volume by volume or the entire array - offing the most efficient and fast coping with almost no downtime.

Another benefit is that after the first copy is made, both storage systems can be written to synchronously to stress-test the new storage system under a production workload. This can be done if needed to gain 100% confidence in the new storage system. Then when the time is right, production can cut over entirely to the new system.

Finally, and most importantly, storage virtualization is agnostic but not equal. It can migrate from any-to-any SAN-based storage system, but beware, as some storage systems require a migration process to preserve existing data. A complete storage virtualization system can see and access existing data from any SAN-based storage system, negating the need to take the storage system offline when

migrating between storage platforms, causing unexpected downtime or a loss of productivity. Simply put, storage virtualization can be the foundational migration component that data centers need. The apparent benefit of selecting storage virtualization as a migration strategy is that it also opens the opportunity to leverage the capabilities of storage virtualization. For example, storage virtualization allows for all storage management and storage services to be provided from a single interface and a single set of tools.

ESF Storage Solutions

Hyperconverged Solutions

Hyperconverged solutions merge compute, storage and networking into a single infrastructure that provides a consolidated single unit containing all necessary functions along with delivering automation and simplifying data center operations. The utilization of virtualization and software-defined storage allows for simplified scaling whereas the new infrastructure can be provisioned on-demand as resources are needed. This eliminates two large challenges faced when dealing with virtual servers, scaling of compute and storage resources.

Hyperconverged environments present one single consolidated unit that is a superset of SDS by the means of utilizing software- defined infrastructure on top of virtualized machines. The virtualized machines are usually standard COTS servers which house storage as well as the compute and completely replaces the traditional storage mechanisms such as LUNs, volumes, and SAN and NAS. The storage capacity from all virtualized machines is used to create a shared storage resource pool available to all servers. This removed dependencies from expensive legacy storage arrays. Not only does this reduce cost and simply scaling, but it requires very little resource management to the extent that management can be done through a single console by an IT generalist. Due to this, HCI is seeing large growth amount enterprise users.

To better understand this, let's look at a few of the powerhouses that dominate the Hyperconverged landscape:

Microsoft Storage Spaces Direct

Storage Spaces Direct (S2D) was introduced with enhancement in Windows Server 2016 that extended the existing software-defined storage (SDS) stack within Microsoft's Server Message Block (SMB) 3.0 protocol. The SMB protocol includes two features, SMB Direct, and SMB Multichannel. SMB Direct implemented the use of various high-speed Remote Direct Memory Access (RDMA) methods to transfer large amounts of data with little CPU intervention. SMB Multichannel allows file servers to use multiple network connections simultaneously and provides fault tolerance through automatic discovery of network paths to dynamically add connections as required. S2D enables a software-defined hyperconverged solution. With the addition of the recently released Windows Server 2019, Microsoft expands the feature set to include deduplication and better monitoring, helping Storage Spaces Direct further.

Nutanix Acropolis

Acropolis is Nutanix's foundational platform for hyperconverged infrastructure that enables virtualization, storage services, virtual networking and cross-hypervisor mobility. Nutanix clams that Acropolis can enable a turnkey solution capable of being deployed in less than 60 minutes. Acropolis allows a single interface for all management including optimization of virtual environments, infrastructure monitoring and allows for automation of everyday operations. Acropolis replaces the need for SAN and NAS-based storage solutions with a cloud model approach to virtualization that offers substantial benefits by simplifying the typical infrastructure life cycle.

Red Hat Ceph

Red Hat Ceph Storage is a leading software-defined, hyperconverged storage solution with the ability to support block, object and file storage requirements. Based on Ceph, the open source storage software that is overwhelmingly preferred by OpenStack users, it has a scale-out cluster design to enable growth of both capacity and performance. There is a front-end (public) network for connecting to clients and a backend (cluster) network for handling storage replication, rebalancing, and recovery. The cluster network can carry up to three times the traffic of the public network for data writes and more when rebuilding or rebalancing nodes. Additional advantages of Ceph include no single point of failure, and software-defined services for self-managing, and self-healing, and to help reduce administration costs.

VMware vSAN

vSAN is an enterprise-class, storage virtualization software that works with VMware's vSphere which virtualizes servers. When the two are combined it allows for a single platform to manage compute and storage. vSAN converts industry standard servers and storage into hyperconverged infrastructure displaying a hybrid cloud model. vSAN eliminates the cost and complexity of traditional storage by providing a unified storage pool, simplifying storage management and deployment. By offering policy-based management vSAN removes the burden of maintaining storage volumes, LUNs and separate arrays. The policies can define characteristics such as capacity, availability and performance and be assigned to various virtual machines, reducing the cost and complexity to improve business agility.

As we look across the hyperconverged landscape, Mellanox ESF with built-in offloads and accelerators enabled delivery of line-rate network and storage performance and improves server efficiencies by optimizing data transfers. A Mellanox ESF offers hyperconverged customers the ability to choose the hyperconverged solution of their choice to deploy storage on low-cost file servers, while delivering the highest performance. This enables storage that rivals costly Fibre Channel SANs with improved efficiency and lower latency and operating cost. For information on configuring RoCE for lossless fabric in VMware ESXi 6.5 and above click here.

HCI Storage Deployments

In the previous section we discussed the top HCI solutions which can be deployed on standard servers which house compute and storage within the same node. In this section we'll expand on HCI companies that have developed a software-based data services platform that can be deployed on your choice of standard x86 servers AND commodity storage media. The advantage here is that these companies free you from proprietary platforms and pricing while still delivering enterprise-class performance.

DataOn

DataON is a leading provider of hyperconverged cluster appliances (HCCA) and storage systems optimized for Microsoft Windows Server environments. Their solutions are built with the single purpose of rapidly and seamlessly deploying Microsoft applications, virtualization, data protection, and hybrid cloud services. The HCCA solution is exclusively focused on customers who have choose Microsoft and therefore the platform they provide is designed for the Microsoft software-defined data center (SDDC) and offer support for NVMe and RoCE. Learn more

Datera

The Datera Data Services Platform (DSP) software combines multiple servers' flash and disk storage and preforms auto-tiering and auto-migration of live data. It is self-configuring and self-provisions the data services that are needed. The software supports quality of service, service-level objectives and multi-tenancy, and since it's software, it can run in a public or private cloud. The DSP software integrates with Docker, Rancher and Kubernetes to deliver storage to containers and provides support for RoCE, NVMe and persistent memory technologies. Together, Datera with a Mellanox ESF provides an uncompromised solution to help businesses realize the value of their data and accelerate them into the future. Learn more

DriveScale

DriveScale's team of entrepreneurs are experts in diverse disciplines who became united by a common quest: to re-imagine static compute infrastructure as adaptable and programmable, making every data center a composable, agile, scale-out cloud. Their solution will give any company the ability to enjoy the same speed, flexibility, and cost-efficiency that the Cloud giants have created for themselves including support for NVMe and RoCE. With the growth of data-driven and distributed applications, enterprise and cloud companies will benefit from the rapid deployment, flexible operations and high availability previously only found in the largest data centers in the world. Learn more

Excelero

Excelero enables enterprises and service providers to design scale-out storage infrastructures leveraging standard servers and high- performance flash storage. With Excelero's NVMesh, which utilizes NVMe and RoCE to build distributed, high-performance Server SAN for mixed application workloads. Customers benefit from the performance of local flash, with the convenience of centralized storage while avoiding proprietary hardware lock-in and reducing the overall storage TCO. The solution has been deployed for hyper-scale Industrial IoT services, machine learning applications and massive-scale simulation visualization. Learn more

EXTEN

Exten's HyperDynamic software disaggregates storage by moving non-volatile memory from traditional storage and server systems to fabric-attached storage appliances or JBOFs (just a bunch of flash) connected using low-latency and high-bandwidth RoCE network for NVMe over Fabric support. The HyperDynamic software runs on a flexible, commodity-based, hardware ecosystem, including industry standard Intel and AMD servers, as well as with custom SoC-based enclosures. These capabilities provide large data centers with the freedom to design highly-optimized solutions using commodity components from diverse manufacturers that balance cost and performance. Learn more

Kaminario

Kaminario pioneered data plane virtualization technology which allows customers to decouple the management and movement of data from the infrastructure it runs on. Changing the relationship between applications and infrastructure changes the economics and agility of business-critical data and lets IT organizations accelerate the movement of business-critical applications to cloud infrastructure by leveraging RoCE and NVMe-oF to unlock the maximum performance. Learn more

Kioxia

Kioxia, formerly Toshiba Memory, offers KumoScale, a shared accelerated storage software that utilizes NVMe-oF and abstracts the storage resources into a virtualized pool. The storage pool can then be efficiently allocated in "right-sized" capacities to compute nodes. The software was developed with a strong focus on virtualizing, managing and securing data. Its feature-rich, standards-based design provides storage management functions for abstraction, provisioning for containers and virtual machines, and integration with popular orchestration frameworks through a secure API. Learn more

Lightbits Labs

Lightbits Labs enables cloud-scale through disaggregation of storage and compute, and transition from inefficient Direct-Attached SSD architecture to a low-latency shared NVMe flash architecture. In stark contrast to other NVMe-oF solutions, the Lightbits NVMe/TCP solution separates storage and compute without touching the network infrastructure or data center clients. With NVMe-over-RoCE,

Lightbits delivers the same IOPS as direct-attached NVMe SSDs with up to a 50% reduction in tail latency. Learn more

MinIO

MinIO is pioneering high performance object storage by leveraging the hard-won knowledge of the web scalers and develop a simple scaling model for object storage and local NVMe performance. At MinIO, scaling starts with a single cluster which can be federated with other MinIO clusters to create a global namespace, spanning multiple data centers if needed. It is one of the reasons that more than half the Fortune 500 runs MinIO. Learn more

Qumulo

As data footprints continue to grow, so do the demands on enterprise storage, revealing limitations in scale, performance, and control of legacy scale-out architectures. The modern design of the Qumulo File Fabric (QF2) provide customers with a truly modern NVMe-oF scale-out storage solution. More economical than legacy storage with leading performance, QF2 is a modern, highly scalable file storage that runs over RoCE in the data center and the public cloud. It provides real-time analytics to let administrators easily manage data no matter how large the footprint or where it's located globally, and continuous replication allows data to move where it's needed when it's needed. Learn more

Scality

Traditional storage and file transfer technologies are just not able to keep up with today's geo distribution and replication demands of cloud storage. Scality solves this problem with an integrated solution, leveraging the Scality RING storage platform addresses the massive storage and data transfer needs of the most demanding companies and enables geographically distributed petabyte-scale storage infrastructures to support geo-distributed workflows, content distribution infrastructures, and active archives. Learn more

SimpliVity

HPE SimpliVity systems provide a software-defined hyperconverged infrastructure that offers enterprise-class performance, data protection, and resiliency. By delivering automated storage utilization, always-on compression, and policy-based VM-centric management, SimpliVity powers the most efficient data centers. SimpliVity delivers a hyperconverged solution that dramatically simplifies IT by combining all infrastructure and advanced data services for virtualized workloads. The data efficiency baked into SimpliVity improves application performance, frees up storage and accelerates local and remote backup and restore functions to deliver extreme data efficiency improvements across all deployments. Learn more

StarWind Software

StarWind provides a unique blend of simplicity, performance, and affordability with flexibility with a low-entry hyperconverged appliances, StarWind also offer software-only or storage appliance variants of its product, therefore being able to adapt to any organizations data center be that entry-level or large enterprise. StarWind's products are designed around a core fundamental of simplicity, performance, and full-features at low cost. StarWind specializes in storage virtualization and building iSCSI, iSER, RoCE, NVMe over Fabrics, Fibre Channel over Ethernet, ATA-over-Ethernet SAN, and NFS and SMB3 NAS on commodity hardware. Learn more

WekaIO

WekaIO offers a shared, fast, parallel file system, WekaFS™, which delivers unmatched performance at any scale compared to legacy storage infrastructures. With the same enterprise features and benefits of traditional storage, Weka tackles the most demanding storage performance challenges in data-intensive technical computing environments so customers can solve today's biggest data-intensive problems that hold back innovation. Optimized for NVMe flash and hybrid cloud, WekaFS accelerates time-to-insight from mountains of data and helps customers get the most out of their high-powered IT investments. Learn more

Zadara Storage

The Zadara Cloud Platform uses a combination of industry-standard hardware and patented Zadara software to deliver enterprise-class speed, security, and scalability — together with the convenience of the cloud. The Zadara Cloud Platform can run multiple tenants, file, block and object storage, simultaneously on the same physical machines. Storage acceleration through RoCE and NVMe, and virtualization and resource isolation, the Zadara Cloud Platform can handle hundreds of nodes while workloads remain isolated from one another. With Zadara your data storage and management system is always in sync with your requirements. Learn more

Dedicated Scale-out Storage Arrays – NVMe-oF

NVMe over Fabrics offers a high-speed communication between compute and storage and delivers new efficiencies in scaling storage infrastructures. NVMe-oF was developed specifically as a storage protocol intended for fast data transfers. It was designed to move data from remote nodes on a fabric as efficiently as internal memory transfers – and proves to have microsecond latencies and scale out to tens of thousands of devices. NVMe-oF offers a replacement for transmitting SCSI over IP or FC networks as both can take advantage of the increased command set and queue depth of NVMe to communicate with NVMe storage. Storage systems with NVMe drives are already available on the market and several of the leaders in this space are listed below.

HPE 3Par

HPE 3PAR all flash storage can now take advantage of storage technology innovations to accelerate application performance with an NVMe Storage Class Memory module and RoCE transportation, which enhances the performance of All-Flash storage systems in specific small block workloads by decreasing the latency for specified volumes. Intelligent caching algorithms are used to extend dram cache to SCM devices through NVMe transport. NVMe SCM module is available as an easy to install add-on card and simply takes advantage of an existing node PCI slot-3. Learn more

Dell EMC

Dell EMC enters the NMVe market with PowerMax, which changes the game for data storage with its multi-controller design, active/active scale-out architecture and industry standard, end-to-end NVMe. PowerMax consolidates block, file, mainframe, and IBM workloads and modern, real-time analytics apps on a single array. The real-time machine learning engine and the use of RoCE automatically optimizes performance with no overhead. Plus, PowerMax offers you the gold standard in replication, 6 9's availability, and data at rest encryption that's FIPS 140-2 validated. Learn more

Pavilion Data

Pavilion offers the industry's first Hyperparallel flash array which is proven in the world's largest NVMe-oF deployments at the most demanding customers who are making the impossible possible. Pavilions Hyperparallel flash array is the world's first end-to-end NVMe-oF storage array. Designed from the ground up to take full advantage of the parallelism of NVMe, and it utilizes a composable/ disaggregated architecture that eliminates the operational, scalability and utilization challenges of DAS. Learn more

NetApp

The NetApp AFF A800 is an end-to-end NVMe array capable of delivering massive performance and comes from NetApp's strong line of AFF arrays. The A800 is aimed at those that have the most demanding workloads needing lots of performance and plenty of storage. NVMe-oF adds a massive performance boost that should appeal to enterprises requiring maximum throughput and microsecond latency for their business-critical applications. The NetApp EF-Series of all-flash arrays deliver fast, consistent response times to accelerate databases, high-performance computing and analytics workloads, offering consistent microsecond response with end-to-end NVMe and industry leading support for 200 Gb on RoCE and InfiniBand. Learn more

HPE Nimble

The Nimble All Flash Arrays (AFA) delivers the fastest and most reliable access to data with performance over 1.2 million IOPS at sub-millisecond latency. The new HPE Nimble systems represent a big step forward because they are fully ready for RoCE, NVMe and Storage Class Memory (SCM) technologies. The new Nimble systems are NVMe-ready in two different ways; First, the chassis can take NVMe drives and the host-side connectivity will be able to use NVMe-oF. The new hardware offers radically increased performance vs the previous product line and for the same money, you get a much faster system. Learn more

Pure Storage

In anticipation of the inevitable shift to NVMe, Pure Storage engineered FlashArray//M to be NVMe-ready from the beginning. Every FlashArray//M ships with dual-ported and hot-pluggable NVMe NV-RAM devices, engineered by Pure. Additionally, the FlashArray//M chassis is wired for both SAS and PCIe/NVMe in every flash module slot and controllers are non-disruptively upgradable to transition internal and external networks from SAS to NVMe. The Purity Operating Environment is optimized for NVMe with massively parallel and multi-threaded design, as well as global flash management across the entire flash pool. As a result, customers will be able to convert any FlashArray//M to NVMe-enabled controllers, utilize RoCE transport and increase capacity without a forklift upgrade or disruptive migration. Pure FlashArray supports NVMe-oF over RoCE and Pure FlashBlade supports NFS and SMB over Ethernet. Learn more

VAST Data

VAST Data's Universal Storage architecture allows for the storage of all data within a single tier of an all-flash storage array that is fast enough for the most demanding applications, scales to exabytes, and is affordable enough to compete with spinning disks, even for inactive data. With costs comparable to HDD, there is no longer a need for tiering of data. Their highly-availability NVMe enclosures manage over one usable PB per RU and can be scaled independent of servers. A scalable, shared-everything cluster can be built by connecting every server and device in the cluster over a Mellanox Ethernet Storage Fabric. Learn more

Summary

We hope this guide will be helpful when considering, designing, and building out your next storage fabric. Sooner is better than later, if application performance is a competitive weapon for your organization and the advantages of an Ethernet Storage Fabric will provide the best business benefits as a high performance, efficient, storage-optimized solution.

To learn more, visit our ESF landing page, or the Mellanox online store to purchase or chat with a sales representative.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. Neither NVIDIA Corporation nor any of its direct or indirect subsidiaries (collectively: “NVIDIA”) make any representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice. Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk. NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs. No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative. liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, and Mellanox are trademarks and/or registered trademarks of Mellanox Technologies Ltd. and/or NVIDIA Corporation in the U.S, and in other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

For the complete and most updated list of Mellanox trademarks, visit http://www.mellanox.com/page/trademarks.

Copyright

PNY Technologies Europe
9 rue Joseph Cugnot, 33708 Mérignac cedex | France
T +33 (0) 5 56 13 75 75 | pnypro@pny.eu
For more information visit: www.pny.eu