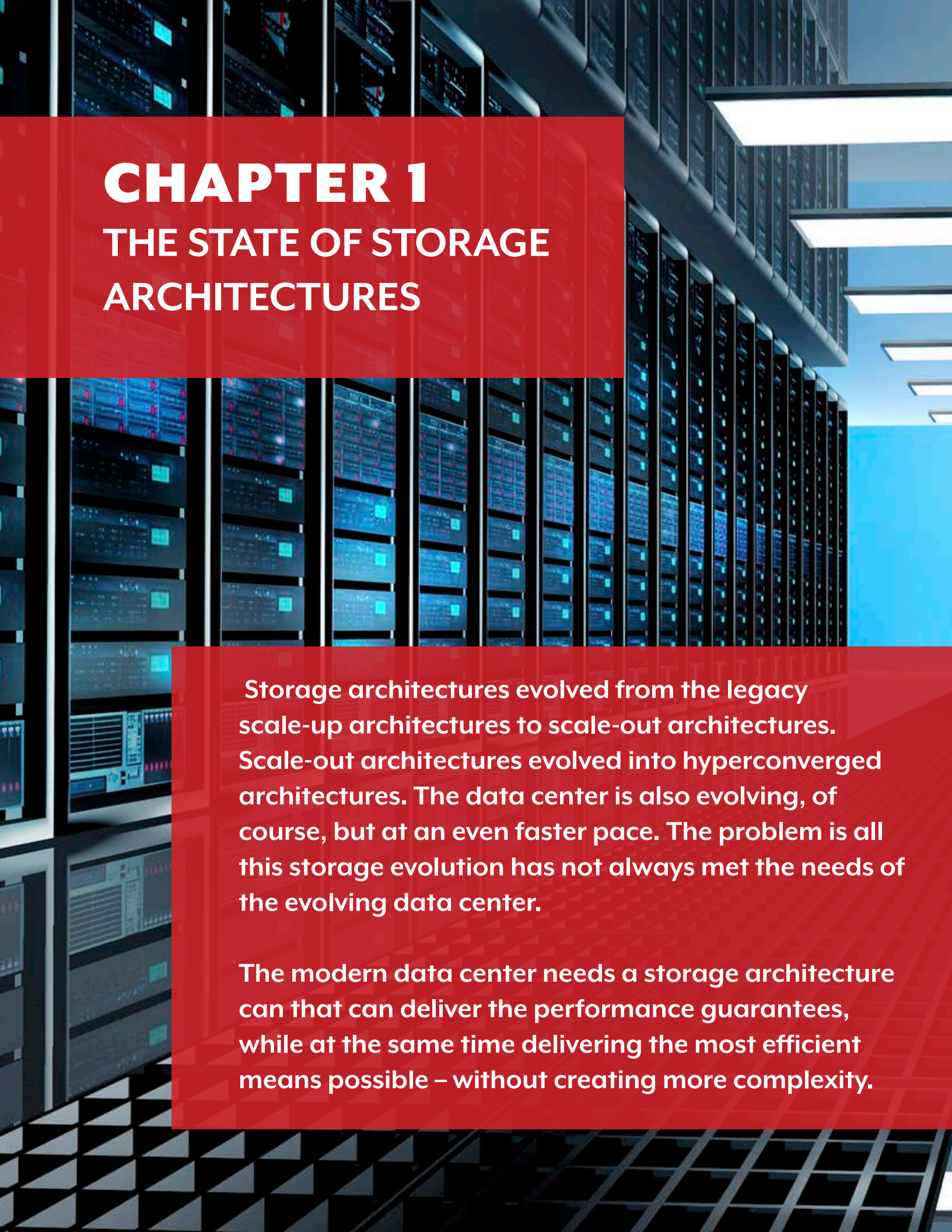
All-Flash Arrays Need A Modern Architecture

by George Crump

Table of Contents



CHAPTER 1
THE STATE OF STORAGE ARCHITECTURES

Storage architectures evolved from the legacy scale-up architectures to scale-out architectures. Scale-out architectures evolved into hyperconverged architectures. The data center is also evolving, of course, but at an even faster pace. The problem is all this storage evolution has not always met the needs of the evolving data center.

The modern data center needs a storage architecture can that can deliver the performance guarantees, while at the same time delivering the most efficient means possible – without creating more complexity.

ARCHITECTURE 1 - SCALE-UP STORAGE

Scale-up storage architectures are the oldest architecture and are based on a fixed number of storage controllers, typically two for redundancy. Shelves, filled with hard disk or flash drives are then attached to those controllers. The number of shelves the controller can support is based on its computing power and IO capabilities.

Workload performance expectation is also a factor in how far the architecture can scale. If the workload is more capacity driven than performance, the impact of overloading the controller with lots of capacity won't matter. But if performance is the concern, each addition of capacity will impact performance and make a difference to how the connecting applications respond.

An advantage of scale-up solutions is they are easy to understand. The design has been around almost since the dawn of the data center. The problem with scale-up architectures is that the organization typically has to overbuy on initial performance so that the system will satisfy future growth. That means for a period of time, the organization is paying for something (performance) that it does not need. Then, at some point, enough workloads are added so the system hits either a performance or capacity limit. That's when IT buys a new system.

Scale-up architectures are ideal to solve a point problem, especially if the workload isn't rapidly requiring additional performance or capacity. They are simple, relatively cost effective and, thanks to flash storage, far more scalable than they used to be.

Flash enables the scale-up architecture to offer a tremendous amount of performance for a considerable period of time. Also, because flash enables a low-impact application of deduplication and compression, these systems can often meet the organization's capacity requirements.

ARCHITECTURE 2 - SCALE-OUT ARCHITECTURES

Scale-out architectures are designed to be the scale-up cure all. These architectures are created from servers, called nodes. Of course a server comes with processing power and then storage capacity is installed inside each node. The nodes are clustered together and storage is aggregated into a single pool. The addition of a node to the cluster automatically provides the cluster with more capacity and compute performance.

The nodes within a scale-out cluster are typically less powerful than the single controller of a scale-up architecture, but the aggregation of the nodes eventually deliver greater performance and capacity. The problem is scale-out architectures typically need a quorum of nodes to start, often three.

The problem is for some data centers, or at least for initial projects the three node requirement, the quorum may deliver more performance and capacity than the organization needs for a long time. In some cases the the organization will never actually need to scale the scale-out architecture.

Another problem for scale-out storage architectures is that they typically need many parallel workloads to reach their potential, they often have limited IOPS capabilities per volume. For example if the organization has one application that needs 1 million IOPS, then scale-up with the appropriate processing power will be a better fit. But if the organization has five workloads that each need a hundred-thousand IOPS and expect to add five more hundred thousand IOPS workloads in the next few years, then a scale-out architecture is a better fit.

SCALE-OUT VARIANTS - SCALE RIGHT AND HYPERCONVERGED

SCALE-RIGHT

To overcome some of the challenges with scale-out architecture, some storage vendors allow their systems to be designated as “scale-right architectures.” These architectures allow the organization to start with a single node used in a scale-up design then scale-out to many nodes when they need. This architecture allows the organization to start small for a particular project then expand the system as IT adds more workloads.

HYPERCONVERGED ARCHITECTURES

Hyperconverged architectures are a form of scale-out storage that instead of requiring dedicated compute leverage the compute of an existing hypervisor cluster like VMware, Hyper-V or a Linux-based hypervisor. The storage software is virtualized and loads as a virtual appliance within the hypervisor cluster.

To some extent, hyperconverged architectures scale automatically since compute, storage and networking components are all added at the same time. For initial implementations this may be ideal, but as the hyperconverged cluster scales problems arise.

First, most data centers don't scale all three vectors (compute, storage, networking) at the same pace. They end up buying more nodes to meet a capacity demand or to meet a compute demand but not both. The result is they end up overbuying on one of the resources.

Second, as the environment scales there are limits on how IT can guarantee specific levels of performance to certain applications. In hyperconverged architectures, everything is shared and ensuring one application gets X number of IOPS is difficult. The only way to make sure performance expectations are met is to build the environment so that all workloads get the same level of performance, essentially overbuying the three resources.

THE NETWORK CHALLENGE OF SCALE-OUT AND SCALE-UP

Both scale-up and scale-out architectures share a problem. The quality of the networking to interconnect these nodes is critical. Often called east-west traffic, this server to server communication is traffic to make sure that all the nodes are in-sync and that data protection levels are met. Hyperconverged exacerbates them because it is also carrying a compute/application responsibility.

THE ALL-FLASH CHALLENGE TO SCALE-OUT ARCHITECTURES

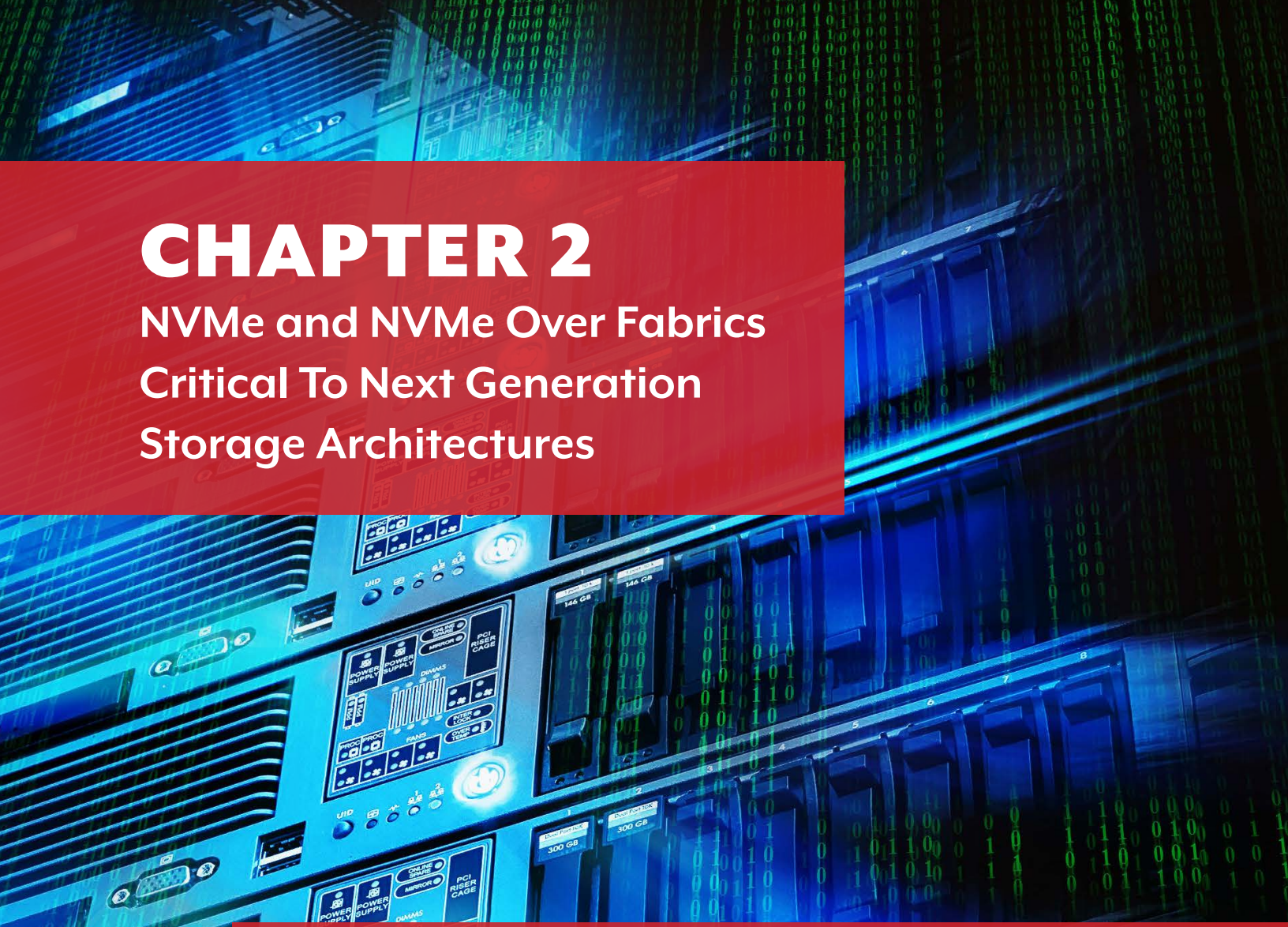
To save cost, most scale-out storage and hyperconverged systems interconnect nodes via basic IP communication. This low level of sophistication was acceptable when scale-out systems where hard disk based. The latency of the hard disk overshadowed the latency of the network communication requirements. But in a flash based system there is no media latency for the network to hide behind. All-Flash scale-out systems also tend to scale further which means more nodes to communicate with and they tend to support more performance critical workloads.

Lastly, all-flash scale-out architectures tend to be more popular in the service provider and very large enterprise. In both use cases there is an added demand for specific performance guarantees, also known as quality of service.

ARCHITECTURE 3.0

The next all-flash architecture is architecture 3.0. This architecture will leverage high-performance, deterministic network infrastructure to interconnect nodes. NVMe over Fabrics may be the only network type that is able to meet that requirement. The architecture will leverage this advanced networking interconnect to also provide the ultimate performance guarantee, virtual, private storage arrays, that can be dedicated to one server or workload.

It will also combine the best aspects of scale-up and scale-out architectures. The system can start as a single node, scale-up architecture and then add nodes to scale performance and capacity. And unlike the legacy scale-up architectures it will not have a per-node or per-volume performance limitations.



CHAPTER 2

NVMe and NVMe Over Fabrics Critical To Next Generation Storage Architectures

The storage media used to be the slowest component within the storage architecture. Now, thanks to flash, it is the fastest. While the performance and low latency of flash allows data centers to make significant steps forward in application scale and response, flash also exposes other weak spots within the storage architecture. IT needs to solve those weak spots in able to fully exploit flash's capabilities.

UNDERSTANDING THE WEAK SPOTS IN THE STORAGE ARCHITECTURE

The storage architecture's weak spots are essentially everything that surrounds the actual flash media, essentially the storage system. The two primary areas of concern are the software that drives the storage and the network within that system. There are two aspects of the network that are particularly important. The internal network that allows the storage software to communicate with the flash drives and the external network that allows the storage system to communicate with either other nodes in the system or to the attaching hosts.

THE INTERNAL NETWORK PROBLEM

Storage systems, whether scale-out or scale-up, are basically servers. As servers they have a certain amount of processing power which, among other things, the storage software uses to move data to and from the flash media. For most legacy storage systems, that communication path is Serial Attached SCSI (SAS). Most flash systems today leverage 12Gbps SAS for that communication. That speed is relatively fast but the communication is still SCSI-based, a communication protocol designed to enable CPUs to communicate with rotating hard disk drives, which had a lot more latency than flash drives do.

A new storage protocol has emerged: NVMe (Non-Volatile Memory Express). It is designed specifically for low latent memory storage devices, like flash. It replaces SCSI and provides a new communications path to memory-based storage. It includes higher queue depth and command counts that take advantage of the low latency flash provides.

As is the case with any new standard, it takes time for it to be adopted and become available. Flash SSD vendors were very quick to start delivering drives that supported the specification, but optimal use also requires the latest PCIe architectures to be grafted into servers. In other words, servers needed to be refreshed. The first implementation of end-to-end NVMe technology is in the latest generation of servers now coming to market.

Storage systems that will first be able to take advantage of NVMe and the performance it provides will be from storage vendors that primarily provide software. They can update their software to directly support NVMe and then load their software directly on these new servers as soon as they become available. By contrast a storage vendor with a system that is less software-defined and more tied to specific hardware platforms will have to wait for their hardware to be refreshed before they can fully exploit NVMe.

THE EXTERNAL NETWORK PROBLEM

The second challenge that flash-based storage systems face is the external network. The external network communicates with other storage nodes in a scale-out storage system and with the physical servers that store and read data.

As scale-out storage systems scale the networking, those systems can become critical. These architectures scale by adding nodes to the cluster. As more nodes are added, the inter-node communication increases. Any overhead in the communications between nodes can be a significant issue for these systems and increase overall system latency.

NVMe is being advanced as a networking protocol, NVMe Over Fabrics (NVMe-F). NVMe-F enables very high speed and very low latency connections. They also typically use some form of remote storage access, minimizing the interaction between multiple CPUs. NVMe-F is an ideal way for scale-out architectures to limit increases in latency as the number of nodes scales.

The final step in optimizing the network component of the storage architecture is the connection to the physical hosts. That connection today is typically either fibre channel or iSCSI-based. While advances in both FC and IP technologies provide the raw bandwidth flash architectures require, they still are burdened with latency and lack of efficiency of a SCSI protocol. NVMe connectivity to the host via NVMe-F will also optimize that communication path. The result should be an eventual end-to-end NVMe communication path that will enable flash to reach its full potential.

"As scale-out storage systems scale the networking, those systems can become critical."

CHAPTER 3
COMPOSABLE STORAGE

The data center needs a new storage architecture. The first architecture, dedicated scale-up storage, provides high performance and efficiency but is operationally complex at scale. The second architecture, scale-out shared everything architectures provides operational simplicity at scale but is less efficient from a compute and storage resources perspective. The time has come for a third storage architecture, Composable Storage – the Storage Architecture 3.0 I referred to in Chapter 1, which delivers the performance and efficiency of scale-up storage with the operational simplicity of scale-out storage.

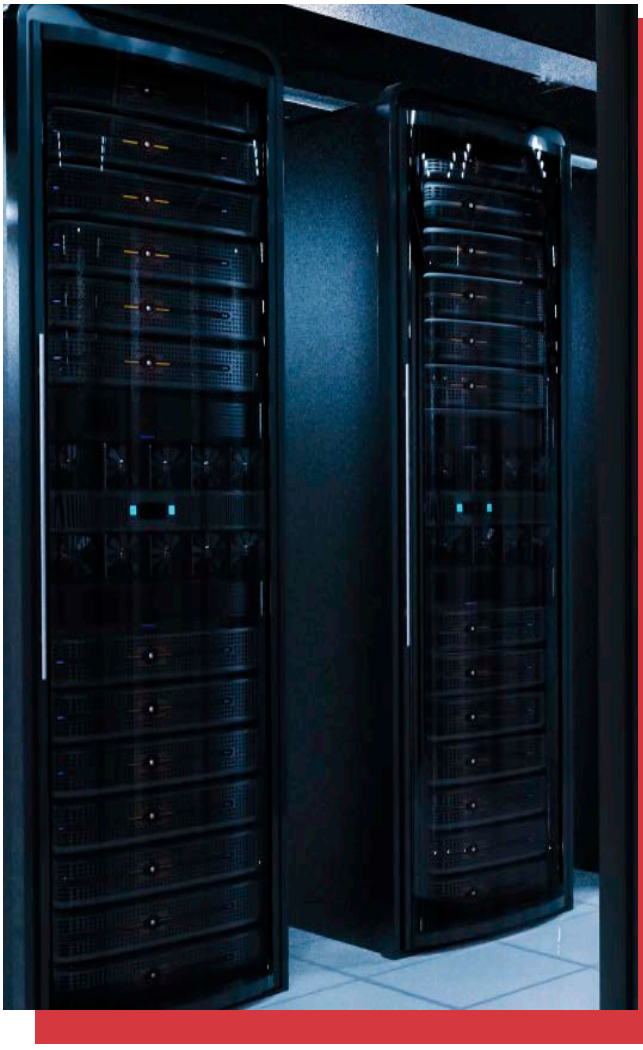
WHAT'S WRONG WITH THE STATUS QUO?

Most data centers have a mixture of applications in the environment. Some are legacy silo'ed applications that require extremely high performance from a finite number of volumes. These applications typically must have very specific guarantees in terms of performance and availability. Scale-up storage systems are ideal for these workloads but as the environment around them grows and the workloads change, it becomes operationally challenging for them to keep up.

These data centers also have a new breed of applications, environments or datasets that are scaling rapidly and are more suited to the scale-out storage design. While performance is important to some of the applications within this group what's more important is the ability to quickly and flexibly scale (in all four directions) to map to the requirements of unpredictable as-a-service application models. Specific consistent guarantees may not be required. Close enough is good enough, in many cases.

Another challenge in scale-out designs is while the compute and storage resources scale well, the inter-node communication often does not. The network that supports the scale-out design, often basic IP, becomes complex and may eventually bottleneck, adding appreciable latency to storage IO.

To deal with the dichotomy, many organizations have multiple storage systems, as many as five or six, with a mixture of architectures, both scale-up and scale-out. There is also a mixture of storage paradigms to solve each particular business challenge – Hyperconverged Infrastructure, Private or Hybrid IaaS, etc. These mixtures makes the storage environment very complex and also brittle.



WHAT IS COMPOSABLE STORAGE?

Composable storage is the third storage architecture. It leverages the best of scale-up and scale-out. Like scale-up architectures, a composable storage system can start with single node. That node can be fully utilized in terms of IO performance and capacity. But, unlike a scale-up design, an additional node can be added to the composable storage so more capacity or compute performance can be available to the environment without introducing another point of management. Composable storage also disaggregates storage compute and storage capacity, which allows for those resources to be dynamically assigned or released by applications using them.

Early iterations of this design were called scale-right architectures. While a vast improvement over scale-up and scale-out architectures, these scale-right designs, once nodes were added, became scale-out and as such inherited many of the negative properties of scale-out. In other words, scale-right really was not a new architecture, merely a bridge between the two existing architectures.

Composable storage, instead being a bridge between scale-up and scale-out architectures is, in fact, an architecture its on right. As it makes the shift from scale-up to scale-out, it addresses the limitations of scale-out architectures. Namely in the ability to dedicate specific performance characteristics to specific applications and it overcomes the potential network bottleneck the internode communication creates as the environment scales.

To address the dedicated performance limitation, composable storage creates dynamically composable virtual private storage system within the storage cluster. This dedicated virtual storage array can be hard allocated specific performance attributes in terms of IOPS, bandwidth and capacity. The virtual storage array can then be used in conjunction with legacy applications were very specific performance requirements are needed.

To address the networking issue, composable storage systems also need to provide better networking. Not only does better networking enable scale, it also allows more complex functions like the virtual private storage array. The problem is advanced networking is expensive and often proprietary. NVMe over Fabrics may give composable storage system vendors a way to deliver advanced networking without being locked into a proprietary or niche networking standard. NMVe enables composable storage to deliver a 4-way scaling capability; scale-up, scale-out, scale-in (less capacity per node) and scale-down (less controllers per cluster).

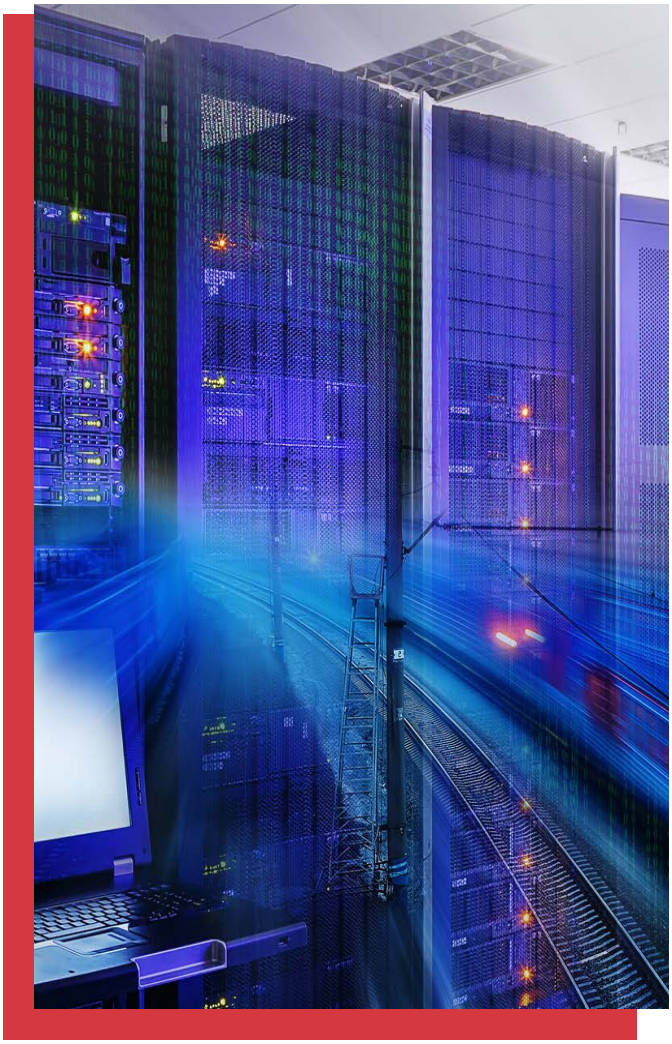
“Scale-right really was not a new architecture, merely a bridge between the two existing architectures.”

NVMe is a new protocol designed specifically for the communication to memory-based storage devices. It is designed to communicate over a PCIe bus and significantly increase command count and IO queue depth. NVMe over Fabrics is the networking of that standard. It enables network performance that rivals a local connection.

Integrating NVMe over Fabrics into the composable storage architectures is a logical step. The nodes within the cluster now communicate performance and latency levels that are almost as low as if they were direct attached. The result is very efficient scale as well as the ability to s cale further.

SOFTWARE DEFINED STORAGE IS KEY

Data centers need the capabilities of composable storage now. They can't wait for storage vendors to design custom hardware and modify their software, especially considering the hardware is available right now. Servers with next generation Intel processors, PCIe buses and full NVMe support are coming to market now. In parallel to the arrival of next generation servers are the arrival of a NVMe flash devices, that promises new lows in latency and new highs in IOPS. In conjunction with the arrival of processors and devices are NVMe over Fabric ready network cards.



If all the hardware components are available, the missing link is the storage software. Software defined storage vendors should be able to quickly adapt their software to the new reality of high performance, NVMe powered hardware and deliver solutions to data centers that greatly reduce the amount of storage systems.

CHAPTER 4

More Than Just Drives - A Complete NVMe All-Flash Solution

All-Flash Arrays bring an unprecedented level of performance to applications in the data center. Most of this performance gain comes from the replacement of hard disk drives with flash as the media of choice. The gain in performance is largely the result of the reduction in latency. But latency is not eliminated, it's just moved. Now, other components of the storage architecture are under pressure to provide similar levels of performance reduction.

CHASING LATENCY

While some latency is introduced by the storage software as it provides more and more features, most latency comes from the interconnectivity of the various architecture components to the flash media. There is latency in the internal connections between the storage system's CPU and the storage software. There is latency in scale-out architectures as they interconnect storage servers (nodes) into the storage cluster and there is latency in the connection to the physical hosts that are attaching to the storage system.

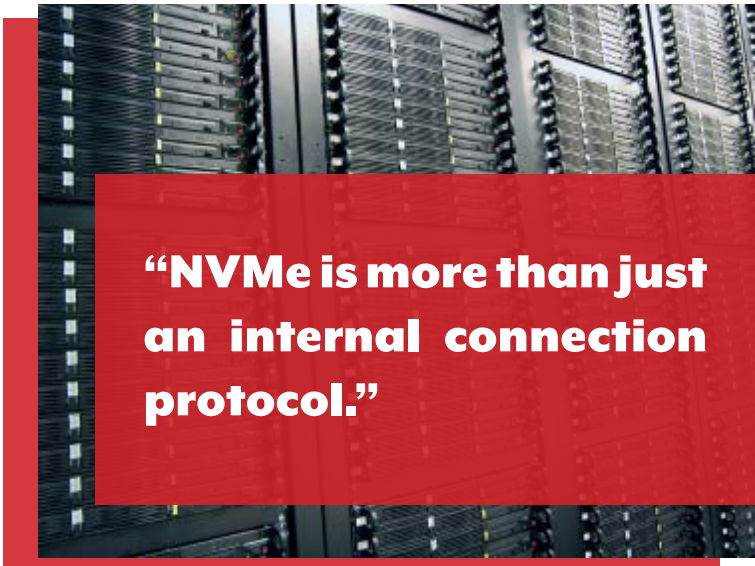
The latency caused by these various interconnections has led some environments to shift to a direct attached storage model only, where the application interfaces directly with internal storage to the server the application is running on. The problem is these applications have to suffer with all the challenges that direct attached storage brings with it, like poor resource efficiency, limited high availability options and difficult data protection integration.

SOLVING THE LATENCY PROBLEM

At the heart of storage architecture latency is the common protocol it uses – SCSI. Introduced in 1986, SCSI was designed for a hard disk era and does deliver the amount of IO commands that a solid state drive can support. As a result, the SSD is actually waiting on the protocol. NVMe was created to solve that problem by drastically increasing IO queue depth and command count. A NVMe SSD provides significantly better performance than a SCSI SSD.

The first step in the NVMe rollout will be for storage systems to use NVMe SSDs and improve the internal communication of the storage server. But NVMe is more than just an internal connection protocol. NVMe over Fabrics (NVMe-F) enables low latency networking outside of the storage server.

The next step will be for storage system vendors to use NVMe as an interconnect between storage servers so scale-out storage architectures can scale without incurring inter-node latency. Finally, NVMe-F will connect to physical servers to deliver shared storage latency that rivals that of internal storage.



"NVMe is more than just an internal connection protocol."

INNOVATIONS FROM A LOW LATENCY NETWORK

While some latency is introduced by the storage software as it provides more and more features, most latency comes from the interconnectivity of the various architecture components to the flash media. There is latency in the internal connections between the storage system's CPU and the storage software. There is latency in scale-out architectures as they interconnect storage servers (nodes) into the storage cluster and there is latency in the connection to the physical hosts that are attaching to the storage system.

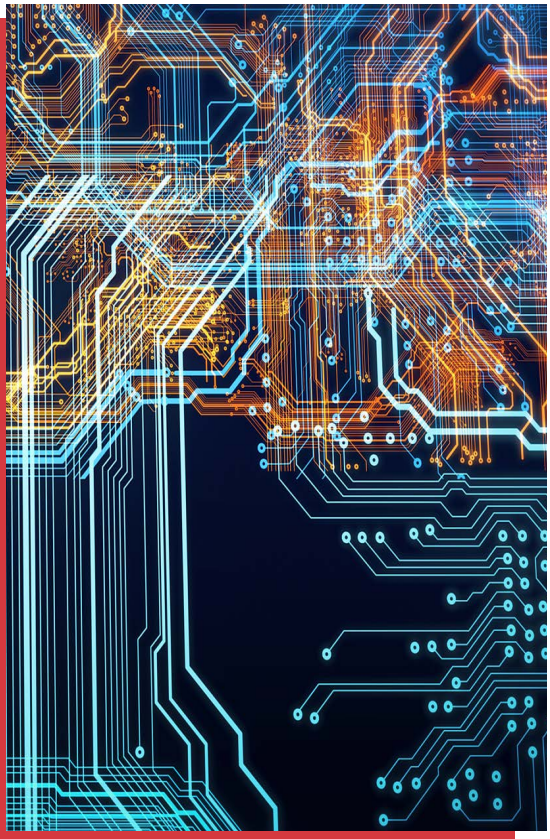
The latency caused by these various interconnections has led some environments to shift to a direct attached storage model only, where the application interfaces directly with internal storage to the server the application is running on. The problem is these applications have to suffer with all the challenges that direct attached storage brings with it, like poor resource efficiency, limited high availability options and difficult data protection integration.

INTRODUCING KAMINARIO K2.N

Kaminario is on its sixth generation of software defined all-flash arrays. The Kaminario K2 Gen6 is a high performance all-flash array built around Kaminario's VisionOS storage operating environment which provides a data service framework and a scale-up and -out architecture.

Capabilities include:

- DataShrink, which provides deduplication, compression and zero detect.
- DataProtect which provides snapshot, replication and encryption.
- DataManage that provides the GUI, Command line interface and a RESTful API.
- DataConnect that provides connectivity with OpenStack, Docker, KVSS, VMware and UCS.
- Kaminario Clarity: a cloud based analytics engine that provides predictive intelligence and support for Kaminario customers.



The K2.N builds on the capabilities of the Gen6 array and will enable customers to take full advantage of NVMe architectures. First, the internal connectivity will be to PCIe-based NVMe drives. Second, the backend connectivity will be fully converged NVMe. This means connectivity between nodes in the scale-out storage cluster will be made via NVMe enabling even lower latencies compared to the already impressive K2 Gen6 which is based on Infiniband.

With the NVMe connectivity Kaminario will also introduce composable storage via Kaminario Flex. The system will be able to create virtual private arrays out of the available storage resources. The administrator will be able to assign a specific set of controllers and capacity to an application or operating environment, assuring application-specific performance. If the solution needs more performance, the administrator can assign additional storage controllers or capacity, all on the fly without disrupting the application. All K2 systems will be able to take advantage of this software orchestration layer.

The rest of the storage architecture like switches and host adaptors will likely convert to NVMe at a much slower pace. To support that, Kaminario will support an open front-end connectivity that can range from fibre channel, to iSCSI and eventually to NVMe. This capability allows the data center to enjoy the benefits of NVMe where it is need most (internal connectivity and internode connectivity) and convert the rest of the environment time and demand requires.

There are parts of the storage architecture that need NVMe right now. While most vendors agree that NVMe SSDs, internal connectivity, is a high priority most are ignoring a second high priority, internode connectivity. Kaminario seems to be the first to address that need. The value of using NVMe for internode connectivity is not just lower latent communications, but the fact that that low latency opens up areas for innovation like composable storage.

“The administrator can assign additional storage controllers or capacity, all on the fly without disrupting the application.”

ABOUT US

STORAGE SWITZERLAND

Storage Switzerland is an analyst firm focused on the storage, virtualization and cloud marketplaces. Our goal is to educate IT Professionals on the various technologies and techniques available to help their applications scale further, perform better and be better protected. The results of this research can be found in the articles, videos, webinars, product analysis and case studies on our website storageswiss.com

KAMINARIO

Kaminario, a leading all-flash storage company, is redefining the future of modern data centers. Their unique solution enables organizations to succeed in today's on-demand wtorld and prepares them to seamlessly handle tomorrow's innovations. Kaminario K2 all-flash array delivers the agility, scalability, performance and economics a data center requires to deal with today's cloud-first, dynamic world and provide real-time data access — anywhere, anytime. Hundreds of customers rely on Kaminario K2 to power their mission critical applications and safeguard their digital ecosystem. Kaminario is headquartered in Needham, Massachusetts, with offices in Israel, London, Seoul and New York City.

THE ANALYST

George Crump is President and Founder of Storage Switzerland. With over 25 years of experience designing storage solutions for data centers across the US, he has seen the birth of such technologies as RAID, NAS and SAN. Prior to founding Storage Switzerland he was CTO at one the nation's largest storage integrators where he was in charge of technology testing, integration and product selection.