GIGAOM RESEARCH

Tangible advantages of SDN-powered scale-out storage

Howard Marks

September 30, 2014

This report is underwritten by Coho Data.

TABLE OF CONTENTS

Executive summary

For decades, most IT organizations have relied on scale-up storage systems for the vast majority of their tier-one applications. Now, however, the modern data center's increasingly demanding workloads combined with – paradoxically – the performance that flash memory provides as a storage medium, have stressed scale-out architecture. Fortunately, a new generation of storage systems has emerged to address the traditional storage system's limitations.

Scale-up storage systems use a small number of storage controllers to manage multiple external shelves of drives. As the performance demands of today's mixed workloads have increased, and flash-based solid-state drives (SSDs) have been increasingly able to meet those demands, the controller has become the bottleneck – the limiting factor in system performance.

In order to have products at various price points, scale-up storage vendors have as many as six models per product line, with each one of them, practically, only scalable over a limited range. Larger models have faster processors with more cores and at high-end proprietary ASICs or other hardware. As users reach the capacity of their systems' controllers, rather than face a painful upgrade process, they order another system. This creates multiple islands of storage.

Instead of two controllers to manage ten dumb shelves of drives, the new generation of storage systems mounts those drives in dozens – or a hundred or more – industry-standard servers and through software running on those servers, creates an integrated storage system. Since the system's compute resources increase with the storage media, scale-out systems increase their performance along with capacity. The most advanced of these systems also take advantage of the latest in local area networking (LAN) technology, software-defined networking (SDN) to optimize the connections from workload servers to storage.

This report will examine the traditional scale-up architecture and the challenges it faces in today's datacenters. It will explore the evolution of the software-defined scale-out approach, discovering where this new model for storage is really a better idea.

Key takeaways:

- Scale-up storage controllers can't deliver the high levels of storage performance that flash-based SSDs promised.

- Technology has surpassed the scale-up architecture, which can no longer handle the demanding workloads in today’s datacenter.
- The new scale-out, software-defined storage platforms have several significant advantages over traditional scale-up systems:
    - They deliver scalability of capacity and performance, on demand
    - They simplify management and provisioning
    - They lower hardware costs through the use of industry standard x86 servers and SSDs
    - They increase performance, data, and system resiliency
- Taking the software-defined, scale-out idea to the extreme produces storage solutions that deliver both performance and capacity while being significantly easier to manage than traditional storage systems.

Situational analysis

Ever since the introduction of fibre-channel storage area networks (SANs) in the 1990s, the vast majority of storage systems have employed the scale-up architecture that uses a small number of controllers (most commonly two) to manage access to a pool of spinning and/or solid-state disks. Additional shelves of disk drives can increase the capacity of these systems, but the range over which any given set of controllers can expand is generally quite limited.

The basic problem is that scale-up storage controllers simply can't deliver the increasingly high levels of storage performance promised by flash-based SSDs. One enterprise SSD can provide enough storage performance to saturate a 10Gbps Ethernet storage network link. However, taking full advantage of that performance requires more compute power and network connections than scale-up systems can provide.

A new generation of storage systems uses software to cluster multiple industry standard x86 servers into a scale-out storage system. Each server adds compute resources and network interfaces to the cluster along with each increment of storage. Scale-out systems can scale over a much broader range of configurations than scale-up systems. At the same time, they can deliver the full performance their flash memory promised.

The latest systems capitalize on the low cost of industry-standard servers while they leverage the latest in SDN technology to improve the efficiency of the cluster – all this while keeping costs down.

The traditional scale-up storage architecture

Scale-up systems use a small number of controllers to manage some number of spinning and/or solid state drives. Midrange, "modular" arrays have two controllers in some variation of an active-active or active-passive configuration. At the high end of the market, scale-up systems can have as many as 16 custom-built controllers with proprietary ultra-low latency interconnects. These systems can achieve higher levels of resiliency than dual controller systems, but only at significant cost for all that custom hardware.

The controllers manage connections from the hosts that consume the system's storage and to the storage in a number of drive shelves that are today connected to the controllers via serial attached SCSI (SAS).

Adding additional shelves of drives can expand the system's capacity, but the controller configuration is generally fixed at the time of purchase.

While adding shelves of drives increases the system's capacity, the limited compute power and cache memory of the controllers must now be spread across more capacity, and presumably more workloads. As a result, owners of scale-up storage systems frequently report that when their systems are expanded, they actually run more slowly.

Enter the SSD

In the spinning-disk era, when a pair of storage controllers was expected to manage a limited number of spinning-disk drives, the back-end disk drives determined the system's performance and the designer of a scale-up storage system could specify the controller configuration to manage as many drives as the back-end SAS connections on the controller could accommodate.

Today's SSDs can each deliver well over 100,000 4K random IOPS or more than 1GBps of bandwidth, enough to saturate a 10Gbps Ethernet port. Even a few SSDs in each drive shelf will easily exceed the ability of the controller to process all the I/O operations they could perform. In fact, just a few SSDs in an external drive shelf can easily saturate the SAS link between that shelf and the controller.

The challenge SSDs bring to storage controllers doesn't end at their prodigious performance. They simply take more work to manage than spinning disks. The controller has to perform system-level wear leveling and other techniques to minimize writes. Add in compute-intensive data-reduction technologies (including compression or de-duplication or both), and even the fastest controllers can only scale so far.

While the traditional scale-up architecture served well in its day, its limitations have become all too apparent:

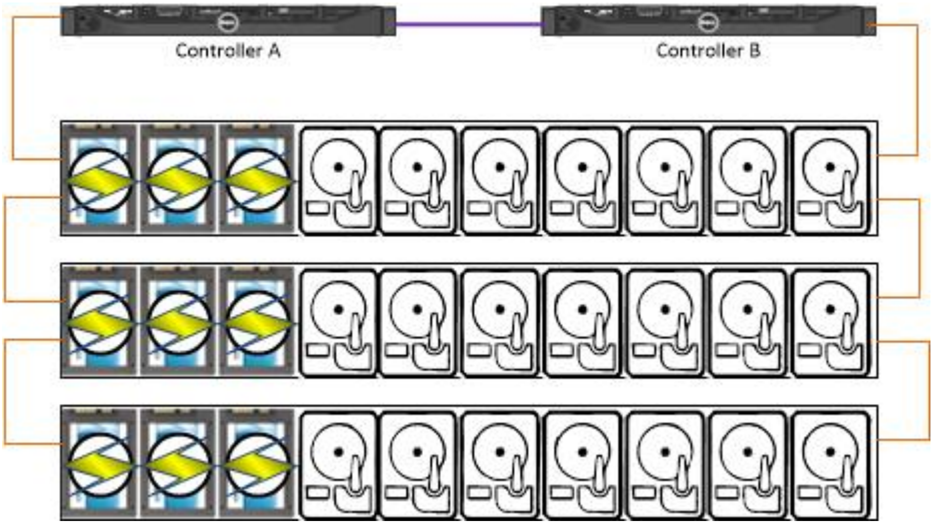
- **Limited scalability**. Since the controller configuration is usually fixed at the time of purchase, the scale-up storage systems, even those that are part of an extended product family like EMC's five model VNX line, can only scale capacity to a limit. In reality most scale-up storage systems have a useful range expansion of 4:1 or less. Because of this limited scalability, IT organizations must manage multiple independent storage systems, reducing utilization – and as a byproduct of having multiple "islands of storage," increasing management effort.

- **Disruptive upgrades**. While adding an additional drive shelf to today's scale-out storage system is usually non-disruptive, scale-up system's scalability is limited – so limited, in fact, that many organizations buy their storage systems fully populated.

  When a customer does outgrow the capacity of its system's storage controllers, the upgrade to a new model usually means shutting the system down, using a more powerful model from the same product family to connect the controllers to the drive shelves, and configuring the new system to deliver the existing volumes. All too frequently, upgrading requires transferring the data from the old to the new system and reconnecting servers to their data volumes. This downtime, or even expectation of downtime, means that migration projects frequently require change-control planning and approvals that can add months to a project, as well as thousands of dollars.

- **Limited fault tolerance**. While scale-out storage systems are designed to survive a controller failure, the designers of dual controller storage systems have a difficult choice. They can limit the system's performance to what that one controller can deliver, typically through an active/passive architecture. Or, they can take advantage of both controllers for performance, but suffer a large performance hit when a controller goes offline.

Figure 1.



Source: Gigaom Research, DeepStorage, LLC

The sad truth is that technology may have just outgrown the scale-up architecture. The old, reliable scale-up controller just can't handle the increasingly demanding workloads, such as virtual desktop infrastructure (VDI), in today's data center. Equally unfortunate, they can't scale to the petabytes of flash that are becoming ever more commonplace in the data center.

Scale-out storage

The controllers have become the performance bottleneck in a traditional scale-up storage system. Scale-out storage systems address this problem by adding not just drives and SSD with each building block, but also by adding intelligence.

Where the scale-up storage system concentrates all the storage management and data services functions in a limited number of controllers, scale-out systems use software to distribute those functions across a cluster of storage nodes. Each storage node has compute power, cache memory, and storage networking ports in addition to some set of drives. Software allows a cluster of storage nodes to act as a single logical storage system.

Architecting each storage node to have its own CPU and storage network port ensures that as high-performance storage devices, like PCIe SSDs, are added to the system, performance does not degrade. Furthermore, sufficient CPU horsepower and network ports are added at the same time. Since, one enterprise SSD can saturate a typical storage port, storage architects who don't maintain the balance between controller CPU, network bandwidth, and SSD performance will find they're not getting as much performance from their expensive SSDs as they expected.

A storage node can be a single controller, typically an industry-standard x86 server, and its local storage or a pair of controllers sharing media as in a scale-out storage system. Shared-nothing systems synchronously replicate data across multiple nodes to protect against node failures whereas systems with dual controller nodes can use a combination of local RAID and replication.

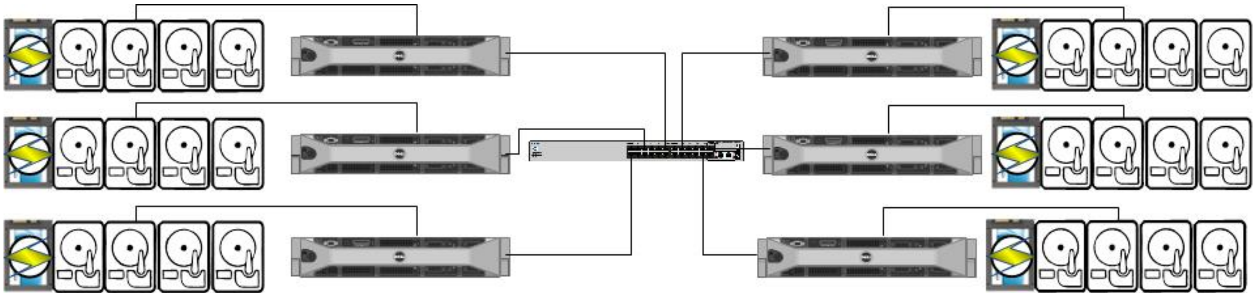
The big advantage of the scale-out architecture is of course the system's ability to scale well beyond the capabilities of a scale-up system. Most scale-out systems are designed to accommodate clusters with as few as two or three nodes to as many as hundreds of nodes. Since the controller workload is distributed across the CPUs of all the nodes in a cluster, a scale-out system can grow from TB to PB simply by adding new nodes to the cluster when they are needed.

This architecture, and the extended scalability it provides, brings several advantages to the modern data center:

- **Greater resiliency.** The distributed nature of a scale-out storage system reduces the impact of device failures in the cluster. A dual-controller scale-up system loses half its controller resources when a controller fails, but an eight-node scale-out cluster with one node failure still has seven-eighths of its resources available.

- **Simplified administration and management.** Since a cluster of scale-out storage nodes forms a single logical storage system, the scale-out system provides single-name space that allows administrators to provision storage from a single pool of capacity across the cluster, freeing them from the overhead of managing "islands of storage."

Figure 2.



Source: Gigaom Research, DeepStorage, LLC.

- **Non-disruptive expansion and upgrades.** In a true scale-out storage system, an administrator wanting to expand capacity and performance only needs to connect a new node to the cluster. New nodes can be configured to provide capacity, performance, or a balance of both.

  Scale-out storage vendors can also take advantage of a system's ability to add nodes to and remove nodes from a cluster dynamically so that it can provide non-disruptive upgrades. When new, presumably more powerful, storage nodes become available, administrators can add them to an existing cluster and take advantage of new technology without having to replace the entire system. Similarly, when an older-generation node reaches the end of its useful life, administrators can eject it from the cluster.

- **Reduced acquisition cost.** Scale-up storage systems are typically purchased with all the capacity they're expected to need over a three-to-five-year useful life. Administrators figure that the incremental cost of additional drives is small and the disruption of an upgrade significant. Since no storage administrator wants to run out of space, most tack on some additional capacity "just in case."

  Savvy scale-out storage users take advantage of their systems' ease of upgrade to apply a more just-in-time (JIT) model to storage purchases. Knowing they can upgrade a system in minutes, after the usual 30-to-90 day purchasing delay, they order less storage initially and expand their systems as demand develops. Since the cost of storage devices decreases over time, by delaying their purchases they can buy a storage node in year three that is 20 percent faster, bigger, and cheaper than the one they would have bought in the initial purchase.

Software-defined storage

Until just a few years ago, delivering enterprise-grade performance and reliability required array vendors to construct custom hardware. Custom ASICs managed RAID-parity calculations, NV-RAM modules with battery, or more recently, capacitor-powered flash-dump power-failure protection. They held the write cache and low-latency interconnects between controllers and helped keep that cache coherent across the controllers. All that custom hardware was expensive to develop and added significantly to the cost of an enterprise storage platform.

As x86 processor performance has increased exponentially – as predicted by Moore's Law – today's x86 systems can achieve in software all the functions that once required custom ASICs. Software-defined storage (SDS) leverages this compute power to provide enterprise-class storage services from the industry standard x86 server parts bin.

A software-defined scale-out storage system can deliver the high reliability enterprise applications expect, despite an x86 server's multiple single points of failure. It does so by distributing the data across multiple storage nodes. These systems can synchronously replicate data between nodes over affordable 10Gbps Ethernet and still deliver latency low enough for even the most critical applications.

Building a scale-out storage system from industry-standard servers has several advantages:

- **Lower cost**. Software-defined storage vendors save costs in two ways. First, they don't have to amortize custom hardware research and development costs across the relatively small number of storage systems they sell. Secondly, competition in the industry-standard server market keeps costs down.

- **Faster adoption of new technology**. Storage vendors typically update their hardware on a more extended schedule than server vendors. These longer development cycles typically leave storage controllers running processors that are generations older than the ones in the latest servers. Software-defined storage developers can qualify existing servers faster than the storage vendors can design and test a newer generation of their controllers. This is especially true for storage systems that are primarily software-based, but include one custom ASIC. Vendors of these systems typically only upgrade their controllers when a new version of their ASIC is ready, thus skipping CPU generations between upgrades. Of course, SSD vendors are also introducing new technologies all the time. SDS vendors can adopt new SSDs more quickly because the SSD and server vendors have already established their mutual hardware compatibility.

Software-defined storage: hyper-converged pros and cons

The most extreme manifestations of the software-defined, scale-out idea are the hyper-converged and server-SAN solutions that run the software controller as a virtual machine or hypervisor kernel process. Hyper-converged systems achieve some economies by using the same compute resources for user workloads and for running the storage system. That economy comes at the price of lower and less-predictable storage performance, as well as a narrow range of scalability.

Many advocates of the hyper-converged model describe a utopian vision of the data center in which every workload can be hosted across an infrastructure of identical, hyper-converged building blocks. In the real world, organizations have varying requirements for compute power and storage resources. If one brick is really all we can use, some customers will be buying storage capacity when they need more compute resources or more CPU horsepower when their users' data fills their existing bricks.

Some hyper-converged infrastructure vendors have partially addressed this issue by offering several different bricks with varying amounts of compute and storage. This allows customers – at least those who have sufficiently characterized their applications – to choose compute-heavy or storage-capacity-heavy bricks carefully. However, the compute-to-storage resources across most vendors' product ranges are limited to three-to-one or four-to-one.

Users simply wanting to add storage capacity must add a full node to the cluster, with all the costs that addition entails, including hypervisor and node management software. These can easily exceed $12,000 for vSphere and its associated tools. Tight hypervisor integration also means that many hyper-converged solutions can only provide storage services for virtualized workloads. Users who have any bare metal workloads will need separate storage systems for different workloads. Since hyper-converged solutions can frequently only provide storage for workloads running in a single hypervisor cluster, hyper-convergence can contribute to the isolated islands of storage we've been trying to eliminate from our data centers for a decade.

SDS storage systems that use dedicated storage nodes can manage all the resources of those nodes to deliver more consistent performance. The designers of these systems can provide each node with sufficient dedicated controller resources so that they can more easily adapt to dynamic application environments. In a hyper-converged system, the storage controller function must compete with the user workloads for CPU resources.

Since these systems use standard storage protocols like NFS and iSCSI, they can provide storage resources for both bare metal and virtualized workloads, including simultaneous support for multiple hypervisors from a single, logical storage pool.

Software-defined networking primer

Traditionally, network devices (including switches and routers) have been independent devices deciding how to filter and forward network packets. These decisions are based on the configuration the network administrator loads into the device and the information the device gleans from the devices it was attached to. While spanning-tree and the various routing protocols allow switches and routers to exchange information about the network topology, each device makes its own packet forwarding decisions.

SDN separates the logical control plane, where decisions about packet forwarding are made, from the data plane that actually moves the data. Rather than distributing the control function across all the devices in the network, software-defined networks concentrate the control plane in a controller, which is frequently a standard x86 server.

Advantages of the SDN approach include:

- **Reduced switching costs**. Since the control plane has been shifted to the controller, switches can use merchant Ethernet switching chips rather the custom ASICs that are typical of conventional switches.

- **More efficient path selection and load balancing**. Since the controller has full visibility into the network, it can make more intelligent decisions about optimal route data flows than protocols based on simplistic factors such as address hashes or number of hops.

- **Reduced management effort and therefore, opex costs**. Conventional networks require that configuration updates be performed on each switch and/or router on the network. By concentrating the control plane in the controller, SDN gives administrators a single point of control, thus reducing the effort required to make changes to the network configuration. More importantly, it eliminates network problems caused by mismatched configurations across multiple devices on the network.

- **Enhanced traffic management via network policies**. Conventional network equipment typically associates functions like access control lists to specific network ports rather than the workloads attached to those ports. An SDN-based network can detect that a virtual server has migrated from one host to another and apply the appropriate policies to that virtual server rather than to the port to which it was attached.

- **Programmability**. Conventional networking equipment determines how it will filter and forward data frames based on a rigid set of protocol rules. SDN technology provides a path for external applications to tell the controller how to set up data flows. Network functions like multi-pathing and load balancing can be performed by an SDN controller and based on richer data sets such as the CPU load on each server.

Combining SDS and SDN

While scale-out system architectures aren't entirely new, the networking part of a scale-out storage design has always been challenging. Not only does the scale-out cluster have to present itself as a single storage system, it also must load balance across storage network ports and nodes. Furthermore, it must replicate data across nodes and find a way to deliver data stored on one node to a server requesting it on another.

The most common scale-out architecture uses a dedicated, high-bandwidth, low-latency network, frequently InfiniBand, to connect the nodes in the cluster. The nodes in a cluster all respond to a common virtual IP address. Connections from hosts to that virtual IP address are assigned to nodes on a round robin or random basis. When a host connected to node D requests data stored on node F, the node the host is connected to (D) requests the data from node F, where the data lives. The data follows the same path back to the host.

The other common model has all the storage nodes replying to a common virtual IP address and using IP redirection to forward requests to the host that wants it. Given the restrictions of classical networking, the connections between hosts and nodes are essentially random.

These traditional scale-out models often result in added latency as the size of the cluster grows. Thus, they have not been used for applications that require high performance storage. The advent of SDN has created the opportunity to change these scale-out models.

An SDN-powered scale-out storage system acts as the centralized controller managing one or more SDN-capable switch. On a conventional network, a host trying to connect to a network file system (NFS) share would send an address resolution protocol (ARP) to find the NFS server's MAC address and connect to the first MAC address it finds. On the SDN system, the virtual interface of the storage system is not just a virtual IP address. It is also a virtual MAC address.

When a host connects to the storage system, the controller can see the whole NFS request and tell the SDN switch to set up a flow from that host to a storage node. Since the controller is itself part of the storage system, it can set up that connection based on information such as the CPU load on a storage node, or which storage nodes hold the most data related to the specific NFS mount point the host is connecting to.

Should a storage node fail, the distributed controller will detect the failure and reconnect the NFS session to another node in the cluster. The switch redirects the data flow to a new node, but the IP and MAC addresses used don’t change and the TCP session doesn’t need to time-out and reconnect. Since NFS lacks the explicit multi-pathing available in iSCSI and Fibre Channel, managing connections in the SDN switching infrastructure makes node failures transparent.

Integrating traditional storage controller logic into the networking fabric itself achieves a new scale-out model for storage that is higher performance and lower cost – both from a hardware and a management viewpoint.

This new storage architecture, which uses SDN to increase the performance, efficiency, and resiliency of scale-out storage, promises to bring ever-higher levels of storage performance to end users, while reducing both the acquisition and long-term management costs of storage.

Key takeaways

- Both traditional scale-up storage system scalability and performance bottleneck at the system's storage controllers. While faster Xeon chipsets have increased the amount of horsepower available to a storage controller, the demands of the modern data center and SSDs ability to deliver that performance have the scale-up model in an uncomfortable set of pincers.

- A new class of storage systems use sophisticated software to combine the power and capacity of a cluster of industry standard x86 servers into a powerful storage system. These systems take advantage of the low cost of today's powerful servers and low-latency 10Gbps Ethernet to deliver high levels of performance and reliability at a low cost.

- Distributing the storage controller's functions across the CPUs of all of the servers allows these systems to scale significantly more broadly than scale-up systems, while reducing the impact of a controller failure. The system's close integration of the storage and network control planes allows the network and storage nodes to interact dynamically. This, in turn, reduces both network traffic and the overhead on the controllers.

- This new class of storage systems is a good fit for today's dynamic data center. Their ability to scale-out on demand and manage network traffic across many storage nodes makes them a good fit for environments like private clouds where agility is a requirement.

About Howard Marks

Howard Marks is a Gigaom Research analyst, as well as the founder and chief scientist at Deepstorage LLC, a storage consultancy and independent test lab concentrating on storage and data center networking. In more than 25 years of consulting, Marks has designed and implemented storage systems, networks, management systems, and internet strategies at organizations that include American Express, J.P. Morgan, Borden Foods, U.S. Tobacco, BBDO Worldwide, Foxwoods Resort Casino, and the State University of New York at Purchase. Testing at DeepStorage Labs is informed by that real-world experience.

Marks has been a frequent contributor to *Network Computing and InformationWeek* since 1999 and since 1990, a speaker at industry conferences including Comnet, PC Expo, Interop, and Microsoft's TechEd. He is the author of *Networking Windows* and co-author of *Windows NT Unleashed* (Sams).

He is co-host, with Ray Lucchesi of the monthly "Greybeards on Storage" podcast where the voices of experience discuss the latest issues in the storage world with industry leaders.

About Gigaom Research

Gigaom Research gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, Gigaom Research addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: research.gigaom.com.

© 2014 Giga Omni Media, Inc. All Rights Reserved.

This publication may be used only as expressly permitted by license from Gigaom and may not be accessed, used, copied, distributed, published, sold, publicly displayed, or otherwise exploited without the express prior written permission of Gigaom. For licensing information, please contact us.