IDC ANALYZE THE FUTURE

IDC TECHNOLOGY SPOTLIGHT

Sponsored by: VAST Data

Legacy architectures have imposed a siloed approach to storage infrastructure because of design limitations. New storage technologies such as NVMe over Fabrics, storage-class memory, and quad-level cell flash media open up the opportunity to transcend these limitations with a new universal storage platform.

Transcending the Limitations of Legacy Storage Architectures Through New Solid State Technologies

August 2020

Written by: Eric Burgener, Research Vice President, Infrastructure Systems, Platforms, and Technologies

Introduction

A perennial challenge in the information technology (IT) industry is how to create the most efficient IT infrastructure while continuing to meet evolving performance, availability, manageability, and cost goals. Over the past decade, vendors and customers alike have been experimenting with software-defined shared nothing infrastructures that promised simpler deployment, expansion, and management, but what has become increasingly clear in the past several years is that these types of systems pose significant challenges when customers try to operate them at scale. As vendors and customers look to deploy the types of IT infrastructure they need to meet the requirements of the digital era, there has been a marked move back toward more disaggregated architectures. The hyperscalers — providers such as Facebook, Amazon, and Microsoft that led the original experiment with shared nothing designs in large-scale web infrastructure — have been very public about their moves back toward disaggregated designs.

AT A GLANCE

WHAT'S IMPORTANT

New solid state storage technologies offer the opportunity to design a highly scalable storage architecture that transcends the limitations of legacy approaches.

KEY TAKEAWAYS

- Enabling solid state storage technologies include NVMe over Fabrics, storage-class memory (3D XPoint), and quad-level cell flash media.
- The resulting disaggregrated, shared everything storage architecture has already changed how leading-edge IT organizations view legacy array designs.

Although shared nothing architectures can be very simple to deploy, expand, and manage, there are issues with them in terms of performance and availability (for certain types of workloads) and efficiency of resource utilization at scale. Shared nothing architectures require complex cache coherency algorithms and can demand a lot of both hardware and data redundancy, which results in inefficient resource utilization (a problem that becomes much more significant and costly at scale). They also require data to be striped across nodes (limiting stripe width and shard) or data reduction metadata to be replicated (limiting data reduction efficiencies) to ensure data recoverability in the event of a node failure. Maintaining cache coherency and dealing with recovery issues in shared nothing environments can take up significant bandwidth because of the volume of east/west traffic, which grows substantially as configurations scale — this is why performance in shared nothing cluster environments does not scale linearly and why recovery times can be a concern in hyperconverged environments.

The disparity between compute performance and network and storage device performance drove many of the storage system design decisions that inevitably resulted in these issues. Historically, compute performance has far outstripped both iSCSI network and storage device performance, particularly when hard disk drives (HDDs) were in use. New technologies, specifically NVMe and NVMe over Fabrics (NVMe-oF), significantly change the performance dynamics in storage and networking and open up the opportunity to reconsider disaggregated storage architectures in a new light.

It is important to note that the architectural potential of this reconsideration goes far beyond the traditional view of disaggregated storage. The NVMe storage protocol enables the use of storage-class memory devices that offer an order of magnitude better performance (in terms of latency and throughput) than SAS-based solid state disks (SSDs) and enable many orders of magnitude higher parallelism (a feature that is particularly important with today's multicore CPUs and the increasing use of big data analytics). NVMe-oF offers similar latency reductions relative to traditional iSCSI-based storage networking. These fundamental improvements don't just mean shared nothing architectures run faster; they open up the option to move to a completely redesigned shared everything architecture, which is much more efficient in terms of resource utilization. They also provide the opportunity to rethink inline data protection, data reduction, and data recovery as well as the applicability of universal storage (i.e., a single platform that simultaneously supports multiple storage access methods) to dense workload consolidation.

System-Level Considerations

Shared Nothing Versus Shared Everything Storage Architectures

Given the performance limitations of legacy SCSI-based storage devices, backplanes, and host interconnects, architects were forced to design different architectures to meet different performance, scalability, and availability requirements. Dual controller system designs could provide low-latency access to direct-attached storage devices, but the controllers themselves limited the performance scalability of the systems. Throughput and bandwidth were limited to what the controllers could process, not the cumulative throughput and/or bandwidth of all the attached storage devices. Maintaining cache coherency in dual controller designs was a challenge, incurring additional latency and complexity as controllers needed to communicate back and forth to ensure data integrity and maintain recoverability as a workload progressed. Multicontroller architectures, while they did provide some additional scalability in these environments, only increased the complexity involved with coordinating all the east/west traffic between controllers. Redundant metadata was often required (at least at some level) in different controllers and/or nodes to maintain cache coherency and data integrity because of the data locality problem. Ultimately, even in these designs, the performance of the controllers limits the scalability of the system (as well as the number of servers to which it can be attached when low latency is required).

A different, scale-out architecture allowed throughput and/or bandwidth to be scaled but could not provide low latency beyond devices directly attached to a given controller (or controller pair) within a cluster node. It also significantly increased the east/west traffic required to maintain cache coherency between cluster nodes, and the fact that this traffic traveled across relatively slow IP-based networks posed another latency problem (whether this was for cache coherency, data services [e.g., global data deduplication], or data access reasons). The relatively slow network performance pushed architects toward shared nothing designs to minimize east/west traffic, but the use of shared nothing designs increased the level of hardware and data redundancy necessary to boost performance, lowered the efficiency of resource utilization, and incurred negative consequences for performance, availability and, ultimately, scalability with larger, more latency-sensitive workloads.

NVMe-oF Technology: A Key Catalyst in Enabling the Move to Disaggregated Architectures

The original design assumptions for shared nothing architectures were based around the fact that it was much faster to access SCSI-based storage devices inside a server than outside a server (across a network). NVMe-oF, a new high-performance storage networking protocol, completely changed that dynamic.

NVMe-oF enables host and cluster interconnects with only a 5 µsec "hop" latency — performance that effectively allows storage to be disaggregated from compute without compromising application performance. This allows much larger systems to be built that can directly connect thousands of servers and tens of thousands of storage devices in a single system in a datacenter without any data locality concerns. A resource on a single server or a single storage device can literally be shared across any of the servers in the configuration without any negative latency implications (from an application's point of view). This allows architects to move to a shared everything architecture — a much more efficient design that needs only a single resilient copy of any resource to share it among all servers and storage.

The implications of a disaggregated, shared everything architecture that does not suffer from the data locality issue are immense. First, it makes for a much simpler, less costly architecture. Any server can access any storage "node" in a clustered configuration with no latency impact. Cache coherency no longer needs to be maintained across different servers, which cuts down significantly on east/west traffic in a storage cluster and eliminates the need for power failure protection hardware that would otherwise be required by volatile and expensive DRAM write-back caches. Without DRAM write caching, there are more consistent access latencies across all the connected storage capacity, and there is zero volatility in the data path. In addition, a disaggregated, shared everything architecture breaks the inflexible association that storage systems have historically exhibited between specific storage devices in a cluster, enabling much more flexible asymmetric scaling that accommodates multiple server types, storage device geometries, and multigenerational hardware technology upgrades. These latter factors support a potentially much longer storage system life cycle, which can realistically approach 10 years, while still enabling access to the latest storage technologies.

Second, it enables a set of consistent data structures for all data and metadata to be globally accessible by all servers attached to a set of shared storage nodes. This global view can be leveraged to implement global algorithms that define how a system builds an atomically consistent namespace and performs global data reduction and data protection in a much more efficient manner than shared nothing architectures. This efficiency directly drives cost savings because it means less infrastructure is required to meet performance, availability, and capacity goals.

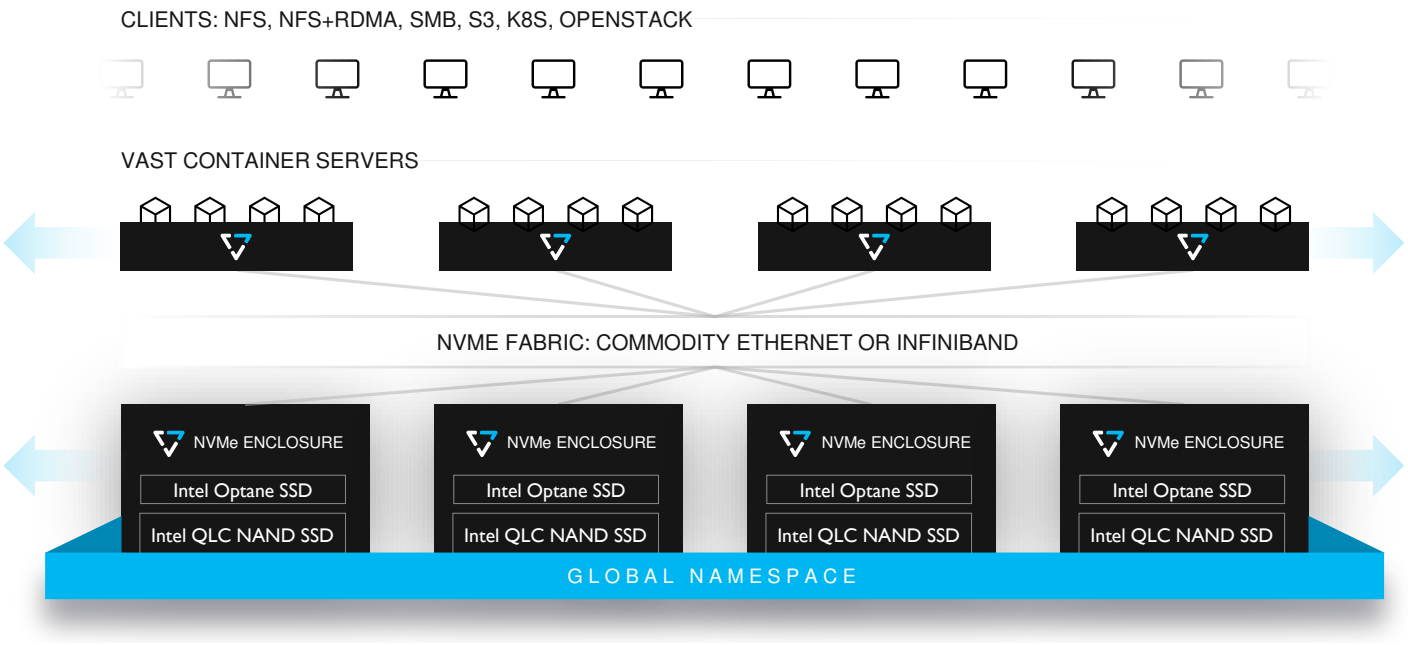
Third, it enables a storage architecture where the metadata about the entire configuration (e.g., the data services, the data) can be kept in the disaggregated storage rather than in the servers (making them "stateless"). When servers are stateless, it is much easier to scale services and fail around any server outage — there is no need to rebuild data when servers fail. Stateless servers better open up the opportunity to use microservices-based, containerized storage operating system designs, abstracting the storage cluster's logic from the underlying server hardware, improving the reliability and deployability of software updates, and providing more flexible deployment options. A stateless server design also makes it easy to dynamically provision server pools to create custom-configured failover domains that isolate traffic and ensure consistent quality of service (in terms of read and write performance) not possible in shared nothing or shared disk architectures. As demand for agility in the IT infrastructure increases, older, more hardware-defined monolithic storage operating system designs are becoming a liability.

Considering VAST Data

VAST Data is a storage start-up headquartered in New York City that delivers a universal storage platform built around NVMe-oF, storage-class memory (NVMe-based Intel Optane SSDs that use a new solid state media type co-invented by Intel and Micron called 3D XPoint), and quad-level cell (QLC) flash media technology. Built on a strong storage DNA that includes team members from Dell EMC (XtremIO), Kaminario (Silk), DDN, IBM, Isilon, Google, NetApp, and Pure Storage, the company has grown first-year revenue faster than the high-water mark set by Pure Storage and achieved unicorn status (i.e., a privately held company with a $1 billion-plus market valuation) in just over a year of revenue shipments. With customers across verticals as diverse as financial services, life sciences, and media and entertainment running a wide variety of unstructured data workloads that include transactional applications, artificial intelligence, machine learning and deep learning, animation, backup, and active archive, the vendor is delivering on the promise of a true universal storage platform with an intelligently thought-out architecture.

Starting with a blank sheet of paper, VAST Data designed a high-performance, highly scalable, and highly agile enterprise storage platform that uses three new storage technologies (NVMe-oF, 3D XPoint, and QLC) in an innovative way to solve the historical problems of dual controller and shared nothing architectures. The use of NVMe-oF let VAST Data design a disaggregated, shared everything (DASE) architecture that allows customers to scale compute and storage resources independently (see Figure 1). All servers are directly attached to all shared storage, and low NVMe-oF latencies ensure there is never a performance penalty for accessing data, regardless of where the data is located in the system. Each attached server runs the VAST Data storage operating system in a container that is easy to deploy and maintain and enables the flexibility to be hosted on a variety of different types of Linux-based x86 servers.

FIGURE 1: ***The VAST Data Cluster Architecture***



Source: VAST Data, 2020

Storage is housed in highly available NVMe storage enclosures that have a mix of Optane and QLC-based SSDs. This mix is part of the vendor's innovative design as the write performance and endurance challenges of QLC are effectively resolved in how the vendor uses the persistent 3D XPoint storage. In this cacheless architecture, all writes are reliably written to the Optane storage, providing extremely low write latencies. Because of the relatively large size of the Optane layer (it is not a tier because the Optane and QLC are managed together as a single tier by the VAST storage OS), writes can be retained for a very long time relative to legacy cache-based architectures before they need to be written out to the lower-cost QLC media. Reads and writes occur from the Optane layer while data access patterns are noticed by the storage OS and, using that data, writes are coalesced into sequential streams of "like data" for eventual destaging to QLC in large block sizes that minimize the need for device- or system-level garbage collection and maximize the endurance of the QLC media.

This approach enables the use of QLC media in write-intensive enterprise environments, and VAST Data provides a 10-year flash media endurance guarantee to underline the viability of this design choice. QLC was chosen to provide most of the storage capacity because of its low $/GB cost, and the blended $/GB cost of VAST Data's Universal Storage Platform, based on the Optane/QLC mix of 3%/97%, is on par with that of nearline SAS HDDs (assuming the 4:1 data reduction ratio the vendor guarantees for mixed workloads). It's interesting to note that QLC read latencies are within 3% of the Optane read latencies, allowing the system to service reads from the entire combined data store at what are effectively Optane latencies.

This design gives a single Universal Storage Platform system the ability to handle the low latencies required by transactional workloads as well as the high degrees of data access concurrency required by artificial intelligence/machine learning/ deep learning and other big data analytics workloads. This flexibility within a single system, combined with the ability to simultaneously support multiple data access methods (NFS, SMB, S3, Kubernetes CSI), makes VAST Data's system an excellent platform for dense mixed enterprise workload consolidation. The vendor's customers use the system in this manner, running a mix of low-latency, mission-critical workloads as well as more cost- and capacity-sensitive tier 2 and other secondary workloads while still meeting their performance, availability, and cost objectives for each.

Because Universal Storage Platforms are designed to consolidate the workloads from multiple existing storage arrays onto a single system, these platforms tend to be large (petabyte scale and beyond). Taking advantage of the DASE architecture, the mix of NVMe-based solid state storage, the lack of data locality concerns, and the scale of deployments, the vendor has redesigned key metadata, locking, data reduction, and data protection algorithms to operate much more efficiently than in traditional storage arrays while providing better performance, higher availability and resiliency, and better resource utilization. Although it is beyond the scope of this paper to discuss these features in detail, IDC would highly recommend that interested customer prospects take the time to understand what the vendor has done with its V-Tree data structure design, its widely distributed Byte-Range Locking, its Locally Decodable Erasure Codes, and its Similarity-Based Data Reduction.

Challenges

Systems built around the limitations of dual controller or shared nothing architectures, caches, tiers, SAS, and HDDs have been able to consolidate some workloads, but they still exhibit issues that preclude the extremely dense consolidation of a wide variety of different types of workloads (even if the configurations were all flash). VAST Data's architecture has resolved this conundrum, but effectively communicating how all the innovations in the platform build upon each other to deliver a compelling value proposition is not easy. It is a challenge to market this solution in a way that clearly communicates why and how it is better than prior, more siloed approaches to enterprise storage infrastructure. VAST Data's rapid sales success in selling to the industry's most sophisticated and leading-edge IT organizations (references are not public but they can be shared under NDA) validates the company's claims, and this success will need to be judiciously utilized as VAST Data generates more awareness around this innovative and ultimately very compelling approach to storage that is well-matched to the needs of businesses in the digital era.

Conclusion

Legacy dual controller and shared nothing storage architectures have significant limitations that have led to designs with performance, scalability, efficiency, and/or cost limitations. A new disaggregated, shared everything approach, enabled by solid state storage technologies such as NVMe-oF, 3D XPoint, and QLC media, promises to overcome these limitations while enabling dense mixed workload consolidation without the performance, availability, recovery, efficiency, or cost disadvantages of legacy designs.

VAST Data's rapid sales success in selling to the industry's most sophisticated IT organizations validates the company's claims for a highly scalable universal storage platform.

VAST Data, a storage start-up that achieved unicorn status in barely more than a year of revenue shipments, offers a Universal Storage Platform based around these three solid state technologies and a DASE architecture that supports mixed unstructured data sets at petabyte-plus scale. The vendor's architecture and well-thought-out design offer compelling value for enterprise customers looking to consolidate file- and object-based workloads onto a single system that can simultaneously meet low latency, high throughput and bandwidth, high availability and fast recovery, efficient resource utilization, and low $/GB cost at massive scale.

About the Analyst

Eric Burgener, *Research Vice President, Infrastructure Systems, Platforms, and Technologies*

Eric Burgener is Research Vice President within IDC's Enterprise Infrastructure practice. Mr. Burgener's core research coverage includes storage systems, software and solutions, quarterly trackers, and end-user research as well as advisory services and consulting programs. Based on his background covering enterprise storage, Mr. Burgener's research includes a particular emphasis on solid state technologies in enterprise storage systems as well as software-defined infrastructure.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
5 Speen Street
Framingham, MA 01701, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com