Vexata VX-OS Architecture

A Technology Review

Release 1.0

Chris M Evans

VEXATA VX-OS ARCHITECTURE

CONTENTS

EXECUTIVE SUMMARY 1

First, Second & Third Generations 1

Vexata VX-OS Architecture 1

BACKGROUND 2

Flash 1.0 2

Flash 2.0 2

Flash 3.0 2

Why NVMe? 2

VEXATA OVERVIEW 3

Architecture Components 3

VX-OS DEEPER DIVE 4

Control and Data Path Separation 4

Distributed Architecture 4

Parallel I/O 5

Product Offerings 5

Scaling Capacity and Performance 5

Enterprise Support 5

Enterprise Features 6

CONCLUSIONS 7

MORE INFORMATION 8

The Author 8

Executive Summary

The demands of the modern data center mean that enterprise storage is on a continuous path of delivering high performance and low latency solutions. Since the introduction of flash to enterprise storage some ten years ago, the evolution of products has seen the adoption of all-flash, reducing costs, and increasingly larger media devices.

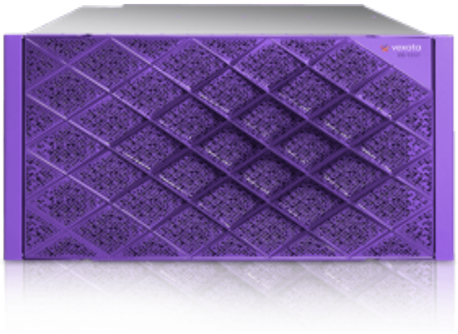
The industry has reached an inflection point where the limits of traditional architectures are being reached. At the media level, the industry has introduced Non-Volatile Memory Express (NVMe) as a replacement for SAS/SATA in an attempt to fully exploit both NAND and storage class memory. NVMe provides very low latency and highly parallel I/O compared to the protocols that were designed in a world of spinning media. As a result, storage system designers have much more bandwidth to play with than ever before.

First, Second & Third Generations

The first generation of all-flash storage systems saw spinning media replaced by all-flash devices. Architectures were tweaked to deal with the issues of endurance, but traditional bottlenecks like back-end SAS still existed. The second generation of all-flash introduced custom-designed products, with features that addressed some of the problems of flash media (like garbage collection), which could impinge on deterministic response times. However, this SSD media was still attached internally using SAS (or SATA) through CPU controller architectures that were designed for spinning disks and further limited performance.

To realise the benefits offered by NVMe, we need new architectures, and this is what the third generation of all-flash arrays will deliver. The third wave of all-flash addresses and corrects bottlenecks seen in previous solutions, fully utilising the capabilities of both NAND and storage class memory (SCM) NVMe devices. Technologies such as Intel Optane can already deliver ultra-low latencies of around 10-20µs, making it essential to make the overhead of a shared storage platform as small as possible.

Vexata VX-OS Architecture



Vexata has developed a new storage architecture built upon the Vexata Operating System (VX-OS) that is designed to fully utilize the performance of persistent NVMe media. The platform was purpose built for modern solid-state media, based on a design that directly addresses issues with previous architectures that restricted scalability or made it impossible to fully utilise new media. The VX-OS architecture includes:

- Independently scalable front-end architecture for host connectivity
- Independently scalable back-end architecture for persistent storage
- Highly-scalable and low latency midplane
- Control and data plane separation with hardware-accelerated I/O path
- Distributed and lockless architecture
- Enterprise feature support

VX-OS is implemented as a distributed software stack that operates within the Vexata VX-100 NVMe arrays. The VX-100 is a hardware platform that adds just 10µs of latency overhead on end-to-end I/O through the array, uses standard Fibre Channel and Gigabit Ethernet connectivity and delivers the storage services, resiliency and availability expected of an enterprise class storage system.

Today, Vexata offers two hardware models, the VX-100F NVMe Flash Array and the VX-100M Memory Class Array. The VX-100F supports up to 435TB of usable capacity, with RAID-5 protection in a 6U chassis. Using standard Fibre Channel interfaces, the VX-100F array delivers a total of 70GB/s of throughput, up to 7 million IOPS at a latency of 200µs. When configured for NVMe over Fabric (NVMe-oF), using 40GbE, the VX-100F throughput increases to 80GB/s with up to 8 million IOPS at a latency of 125µs. The VX-100M is based on Intel Optane, with a maximum capacity of 32TB, delivering 80GB/s and 8 million IOPS at 40µs latency (Fibre Channel).

Without new storage architectures, the price/performance improvements offered by NVMe will be unrealised. Vexata has developed a solution designed to scale with the capabilities of new media and to exploit those new technologies as the industry delivers larger capacity products and matures SCM.

Background

The modern business thrives on data. The sheer volume of information created by companies has been driven by the adoption of mobile technologies and the ability to collect data from a range of devices that collectively are described as the Internet of Things (IoT). Of course, collecting data isn't enough; this information has to be processed and analysed to extract value in real-time. This is very evident when you consider use cases such as machine learning for artificial intelligence, where neural networks are being built upon very dense compute cores and GPUs. These solutions are driving business outcomes based on real-time analytics against very large data sets. Unfortunately, until now, networked storage systems have not advanced to match the capabilities of the deep learning compute processing that are at the core of emerging autonomous systems.

As processing power has accelerated, so the demands on persistent and shared storage have increased. The adoption of solid-state media in the form of flash storage has introduced greater performance and faster response times than ever before. However, even the initial all-flash solutions are showing their age. We have moved through three distinct phases of flash technology, each of which as addressed performance issues in the data center.

Flash 1.0

The first phase was characterised by solutions that simply retro-fitted existing arrays with flash technology. To begin with, these systems were not all-flash, but had tiers of flash storage that targeted the most active I/O in the application. The aim here was to address performance bottlenecks and speed up applications, with targeted deployment of what was an expensive resource. From a technology perspective, NAND flash SLC-based storage was quickly superseded by cheaper MLC and eMLC devices.

Flash 2.0

The second wave of flash saw the introduction of custom solutions, specifically designed to work with flash media. Internal design constraints within NAND flash (such as endurance) can impact the delivery of deterministic response times, so generation 2 products looked to resolve these issues. The second generation of flash products quickly saw a race to the bottom in terms of cost, with new media such as TLC driving price reductions and increasing system capacities.

Flash 3.0

As we move to the next generation of solutions, the media itself is changing. The constant drive to optimise performance and reduce latency means the overhead of 40-year old protocols are starting to show. SCSI, the protocol driving today's Fibre Channel and SAS drives has started to prove inefficient. ATAPI, which underlies SATA has similar problems. These protocols were designed in an era of hard drives and response times measured in milliseconds. Today, when describing flash technology, we measure in timescales a thousand times faster.

Flash 3.0 introduces new architectures specifically designed to work with ultra-fast media like NVMe SSD and Storage Class Memory (SCM). The NVMe protocol removes the inefficiencies of SAS/SATA, while at the same time exposing other parts of the architecture to bottlenecks. In the same way that storage vendors had to design for flash, the current generation of products need to be designed for NVMe.

Why NVMe?

To understand why the NVMe protocol was needed, we have to look at how performance of the whole technology stack has improved. Flash media has I/O latency around 100µs, with storage class memory such as Intel Optane delivering figures as low as 12µs. Traditional SAS/SATA protocols were designed for spinning media and simply introduce too much latency into the data path and fail to capitalise on the parallel nature of reading and writing to solid-state media.

NVMe eliminates the protocol bottlenecks and provides the capability to deliver much more I/O parallelism than was ever achieved with either spinning disk or the first wave of flash-based SSDs.

Vexata Overview

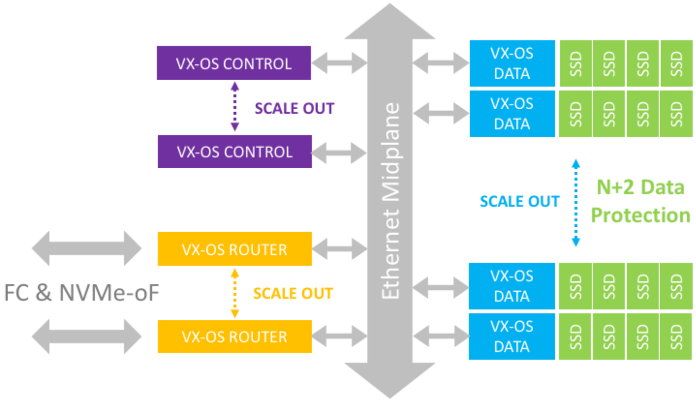
Vexata was founded in 2014 by a team of industry veterans from EMC, VMware, Brocade and others on the premise that every business will be challenged to deliver data intensive applications in real-time. The founding team set out to deliver a shared storage system that can exploit the new wave of persistent storage media in a way that didn't compromise performance, scale or cost. The company initially built the core software architecture (called VX-OS) and a family of products (called the VX-100) that deliver extremely high throughput at very low response times.

What makes VX-OS different from other vendor approaches is that it looks to eliminate the traditional bottlenecks of shared storage while exploiting the features of NVMe. As a result, the VX-100 series of solid-state storage offerings are able to deliver market-leading levels of I/O performance in only 6U of rack space and using a fraction of the memory and compute resources of other vendors.

Architecture Components

The Vexata platform and architecture is based around three major hardware components. At the front-end of the system, I/O controller modules (IOCs) manage host connectivity and data services. At the back-end of the system, persistent storage is provided through Enterprise Storage Modules (ESMs). Both IOCs and ESMs are connected together through an Ethernet midplane.

Today, although Vexata's hardware offerings are based on a dual active/active controller design, the architecture supports scale-out through multiple IOCs and ESMs. This can enable scaling of both back-end performance and capacity, while extending front-end connectivity and throughput.



VX-OS ARCHITECTURE

Key features of the architecture include:

- **Control & Data Path Separation** – the IOCs separate control and data plane functionality, reducing the impact on latency and performance that would normally be experienced in a dual controller architecture. Each IOC implements multiple FPGAs, directly connecting hosts through the Ethernet midplane to ESMs. IOCs are stateless, allowing front-end connectivity and performance to be scaled if necessary.

- **Distributed Architecture** – the traditional functions expected of a shared storage architecture have been distributed across the Vexata solution. Metadata is retained in memory and distributed across both ESMs and IOCs. The architecture itself is lockless. This provides scalability in terms of metadata management and consequently the number of objects and volume of storage that can be stored in a single system.

- **Parallel I/O** – Separation of the control and data planes allows for much greater parallelism in I/O. One of the key benefits in using NVMe is the ability to write many parallel streams of data to each NVMe device, so effective storage architectures need to be able to exploit this capability. VX-OS is designed with parallel I/O in mind and removes the bottlenecks seen in traditional controller-based architectures.

- **Enterprise Support** – Vexata VX-100 systems are designed to work with existing enterprise data center environments. This means offering support for both current Fibre Channel and imminent NVMe-oF protocols with seamless upgrades to protect customer investments in the core platform. Customers can be confident in being able to implement the Vexata solution, without rip and replace, while transitioning to faster protocols in the future. Enterprise support also means providing data services such as snapshots and clones, which are natively supported in the architecture.

End to end, the VX-OS storage operating system software introduces less than 10µs of latency into the I/O path, with around 25-200µs of latency attributed to the storage media.

VX-OS Deeper Dive

As we dig deeper into the design choices made by Vexata, we will expand on the features already discussed.

Control and Data Path Separation

With the introduction of NVMe SSD and storage-class memory, reducing the length of the I/O path from host to media becomes critically important. In traditional architectures, the introduction of data services, such as compression, de-duplication, snapshots and data protection result in elongated code paths and increased latency. This is because these features are delivered by the same processor and memory used to move I/O from front-end host connections to back-end media.

The VX-OS architecture separates control and data planes at the front-end of the architecture using two components called VX-Control and VX-Router. VX-Control delivers platform management functionality, exposed through GUI, CLI and RESTful API interfaces. This component is also responsible for snapshot management, encryption key management, volume management and delivering non-disruptive code upgrades.

VX-OS Router implements a direct hardware path from front-end connections to persistent media. This provides the benefit of both hardware-accelerated assist for data services and removes the bottleneck of shared processor and memory. VX-Router and VX-Control can be independently scaled if necessary.

Figure 1 - VX-100 Rear View - Controllers

The VX-OS Router also manages data rebuilds in the case of an ESM failure. Rebuilds are offloaded to the onboard FPGAs, allowing rebuilds to easily reach 2TB per hour. Encryption is implemented at rest within the routers to AES-256-XTS or AES-256-CBC standards. Key management is performed within VX-OS Control, storing keys either on the array or using standard external frameworks.

Snapshots are performed by VX-OS Control, although intensive tasks such as snapshot consolidation are offloaded to VX-OS Router (more on this later).

At the back end of the architecture, ESMs provide the connectivity to persistent media. Each ESM contains a component called VX-Data and four NVMe SSDs (either NAND flash or Intel Optane SCM). The ESM uses a MIPS-based multi-core processor to manage media metadata and perform the translation between Ethernet and PCI Express connecting the flash or Optane drives.

Distributed Architecture

VX-OS uses a distributed architecture in order to gain scalability with consistent performance. VX-Router and VX-Control can be scaled independently from each other and from other components of the architecture. VX-Data on the ESMs holds metadata on each of the drives and their volume mappings, making it easy to scale capacity and performance simply by adding more ESMs.

The distributed memory on each ESM is matched to the storage capacity it supports on the connected NVMe drives. This removes any constraint on scalability typically seen in architectures that use shared processors and memory. In some existing architectures, the ability to scale capacity has been directly affected by the amount of memory available in each controller node. Memory capacity in the Intel x86 architecture is directly related to the number of processor sockets and so scaling past a certain point means adding more nodes or processor sockets, even if the processor cores never get fully utilised.

Parallel I/O

The final component in the distributed architecture of VX-OS is using a redundant Ethernet midplane. With a redundant midplane, any IOC can talk to any ESM. This means there is no constraint on the amount of I/O that can be directed to persistent storage, other than the physical connectivity of each ESM. Contrast this to an architecture where each drive is associated with only one or two controllers. The ability to drive full use out of an SSD in that scenario is dependent on the capability of the controller.

When any host can talk to any piece of media, it becomes possible to implement a highly scalable parallel architecture. Without NVMe, however, this architecture wouldn't be able to operate effectively. The NVMe protocol supports up to 65535 I/O queues of 65535 requests each, per drive. As flash drives continue to scale in capacity, the ability to write more I/Os to each device in parallel will become more important.

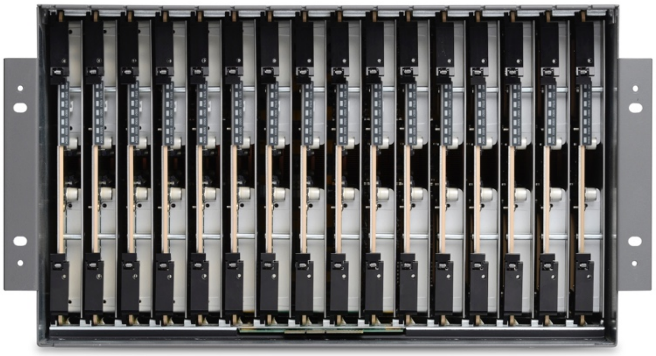


Figure 2 - Front View, 3-16 ESMs

Product Offerings

Today Vexata offers two hardware solutions based on the VX-OS software architecture, the VX-100F and VX-100M. The VX-100F is based on NVMe NAND flash storage, whereas the VX-100M uses Intel 3D-Xpoint (Optane). VX-100F offers high performance and low latency in a scalable platform, whereas VX-100M delivers ultra-low latency with reduced storage capacities. This is reflective of the current media available in the market today, rather than any specific design constraints.

The performance characteristics of both VX-100F and VX-100M are shown in "Table 1 - VX-100 Features". Both Vexata VX-100 platforms are built on a 6U chassis. Each system can be deployed with 3 to 16 Enterprise Storage Modules (ESMs) and two IOCs (I/O Controllers), all of which are hot-pluggable.

Table 1 - VX-100 Features

Feature	VX-100F (FC)	VX-100M (FC)
Media Type	NVMe NAND Flash	NVMe Intel Optane
System Throughput (random 80/20 read/write)	7 million IOPS	7 million IOPS
Typical Latency (random 80/20 read/write)	220µs	40µs
Bandwidth	70GB/s read/write	80GB/s read/write
Maximum Capacity (Usable)	435TB (8TB media)	30TB (750GB media)

Scaling Capacity and Performance

Vexata has built the VX-100 solution to allow for greater capacity scaling while maintaining the performance profile at scale. Today this means support for 8TB flash drives, providing 25TB of usable capacity per ESM and a total system capacity of 435TB using RAID-5 protection or 406TB with RAID-6 (once spares are taken into account). This is based on configurations of either 15D+1P or 14D+2P.

Enterprise Support

VX-OS is designed for the enterprise. Today, the most common protocol in storage area networking in the enterprise is Fibre Channel. IT organisations have depended on the reliability of Fibre Channel, as well as the scalability offered by the latest Gen5 (16Gb/s) and Gen6 (32Gb/s) hardware. In the future, Fibre Channel will move to include support for FC-NVMe, where the storage protocol is NVMe rather than SCSI. This is an important aspect because it allows customers to deploy NVMe based storage into an existing fabric without requiring any changes to the application stack. Both traditional FC and FC-NVMe can co-exist on the same infrastructure.

For companies heavily invested in Fibre Channel infrastructure, such as switching and cabling, this represents the opportunity to see gains in performance, without ripping and replacing large parts of the SAN infrastructure.

Where even greater levels of performance are needed, NVMe-oF (NVMe over Fabrics) will provide faster connectivity using initially 40Gb then 100Gb Ethernet. With the release of VX-OS 3.5, Vexata hardware will support 16 ports of 40Gb Ethernet and a mixed configuration of eight 40GbE and eight 32x Fibre Channel to provide customers with the ability to support mixed modes of operations or to provide a migration path from SCSI based implementations to NVMe over fabric implementations.

Enterprise Features

Enterprise customers expect features within a storage platform that can support their business processes. Snapshots and clones are used to provide replicas for data protection and ongoing production operations. A snapshot may provide a point-in-time copy to recover an application, whereas a clone could be used to build out a test or parallel production environment.

VX-OS implements both snapshots (read-only) and clones (read-write). Snapshots use a redirect-on-write technique that doesn't require any data copying during the snapshot process. Instead, the snapshot is created by manipulating metadata stored in memory, which means the process occurs extremely quickly and with no impact on application performance.

Snapshots are immutable whereas clones provide the ability to update a replicated volume. Both copy processes are implemented in a highly space-efficient way.

The efficiency is achieved because initially, a clone/snapshot and the source volume both point to the same back-end data on flash. Over time, as data is updated, the copies diverge, with fewer common blocks.

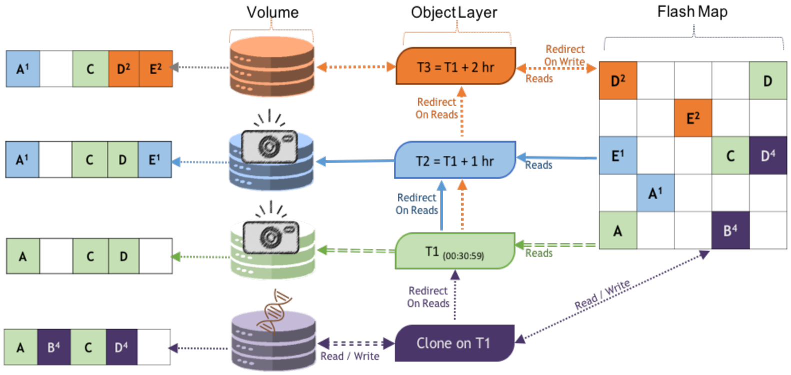


Figure 3 - VX-OS Snapshot Process

Snapshot deletion can prove a significant overhead for traditional storage arrays, as the mapping of volumes that share back-end storage has to be reconciled against snapshot/clone metadata. VX-OS manages snapshots through VX-Control, with the heavy lifting work of consolidation handed off to VX-Router. Today VX-OS supports a maximum of 255 snapshots per volume and a total volume, snapshot and clone count per system of 65472.

As previously mentioned, encryption uses AES-256-XTS or AES-256-CBC standards. Data at rest encryption (DARE) is configured when data groups are defined. The encryption process is hardware accelerated in the data path and so incurs no performance penalty.

Key management is automated when using the Local Key Manager (LKM) feature of VX-Control. LKM automates the generation of Authentication, Data and Data Encryption keys which are then fragmented and further encrypted before being dispersed across the ESMs.

Accelerated encryption in hardware means ESMs don't need to depend on Self Encrypting Drives.

Conclusions

Vexata has created a platform custom-built for NVMe flash and SCM media, without the encumbrances of legacy architectures. This positions the platform to fully support future increases in flash capacities as well as new faster SCM products as this technology develops.

More important perhaps for enterprise customers is the ability to implement VX-100 systems without having to deploy lots of new infrastructure. VX-100 arrays can be slotted into existing Fibre Channel SANs and will seamlessly migrate to new storage networking as the standards fully develop and mature.

VX-100 systems will deliver benefits in new application requirements such as machine learning and artificial intelligence, where consistent low latency and high throughput are essential. These are requirements that simply can't be met by legacy all-flash designs.

More Information

Vexata VX-OS Architecture Overview First Edition

Published by Brookend Limited.

Document reference number BRKWP0107.

No guarantees or warranties are provided regarding the accuracy, reliability or usability of any information contained within this document and readers are recommended to validate any statements or other representations made for validity.

For additional technical background or other advice on replication technologies, contact enquiries@brookend.com for more information. Architecting IT is a brand name of Brookend Ltd and independent consultancy, working for the business value to the end customer.

Email: architectingit@brookend.com
Twitter: @architectingit

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform. After working abroad, he co-founded an Internet-based music distribution company during the .com era, returning to consultancy in the new millennium. Chris writes a popular blog at http://blog.architecting.it, attends many conferences and invitation-only events and can be found providing regular industry contributions through Twitter (@chrismevans) and other social media outlets.