White Paper

Data Protection over NVMe Media

August 2018

a MICROCHIP company

Contents

About this Document

As NVMe SSDs become broadly adopted as primary storage in the enterprise, data availability and ease of operationalization become paramount. Primary storage applications looking to realize the IOPs and latency benefits unlocked by the NVMe storage interface are left with two choices: either software data protection, which consumes CPU resources, or a traditional RAID controller that, left unchanged from traditional architectures, may also penalize performance due to the nature of its proven, but legacy, architecture. What if a new RAID controller architecture emerged that keeps all the traditional benefits while preserving the performance and latency advantages enabled by NVMe drives? In the following pages, a review of architectural concepts and a thesis offer proof that hardware RAID controllers have a bright future in the new NVMe-centric data center.

Definitions

Terms	Definition
ASIC	Application-specific integrated circuit
CPU	Central processing unit
I/O	Input-output transaction
IOPs	Input-output operations per second
NVMe	NVM Express or Non-Volatile Memory Host Controller Interface Specification
PCIe	Peripheral Component Interconnect Express
RAID	Redundant array of independent disks
SCSI	Small computer system interface
VFS	Virtual File System

Trends in Primary Data Enterprise Storage

End users continue to rely on RAID for protection of primary storage. Crucial applications of RAID include protection from drive failure for boot volumes, primary data storage, and database applications.

The cost points of NVMe devices have mostly confined NVMe use to high-performance applications such as caching and journaling. As cost reductions and consolidation occur, it is anticipated that enterprises will increasingly adopt NVMe devices for primary storage, in particular when per-drive densities of less than 4 TB are required (IDC).

In addition, the following ecosystem barriers for NVMe as primary storage are being removed.

- Enterprise reliability (write endurance)
- Serviceability (hot plug, surprise plug)
- Standardization (SFF TA-1005 “Specification for Universal Backplane Management” and SFF TA-1001 “Specification for Universal x4 Link Definition for SFF-8639”)
- Technology for programmable connectivity of SAS/SATA/NVMe (tri-mode)

The removal of ecosystem barriers and the rising popularity of NVMe drives in the enterprise server space mean that RAID for NVMe SSDs will be a mandatory portfolio offering by the end of the decade.

Options for RAID in NVMe Ecosystems

The challenge with applying data protection to NVMe ecosystems is in retaining the performance benefits offered by the NVMe drives without unduly burdening the host system. System designers are typically left with two options:

Option 1: Software RAID using the host CPU

- Uses CPU/chipset resources to provide parity and redundancy
- Direct communication from the host to the SSD, possibly through a PCIe switch

Option 2: Traditional controller-based hardware RAID

- Uses offload capabilities of an ASIC to generate parity and redundancy
- Indirect communication with stages of protocol translation

Both approaches have advantages and disadvantages. A summary of today’s architectures is provided below. There is a third option, however—an evolution of controller-based hardware RAID that is better suited for NVMe.

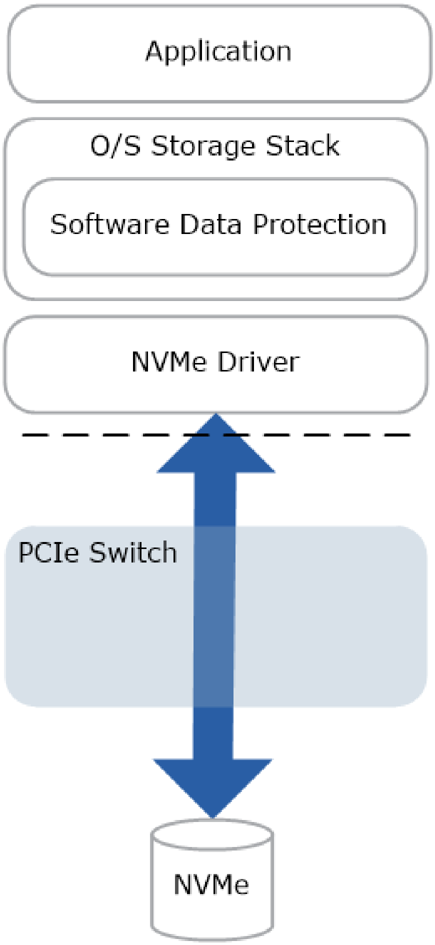
- Combining a PCIe switch with a traditional controller to provide two paths to the drives – one path through the controller’s hardware engines that is intelligently used when offloads are needed and a second path through the PCIe switch that provides the lowest latency when hardware offloading is not needed
- Using a multi-path aware driver to select the correct path for each operation

Data resiliency with high IOPs, low latency, and the benefits of hardware offload are possible for PCIe-attached drives with the appropriate architecture.

Option 1: Traditional Software RAID

Software RAID utilizes the in-box NVMe driver to access PCIe-attached NVMe drives directly or via a switch. A detriment of software RAID is the consumption of expensive compute and memory resources on the host for operations such as the generation of parity.

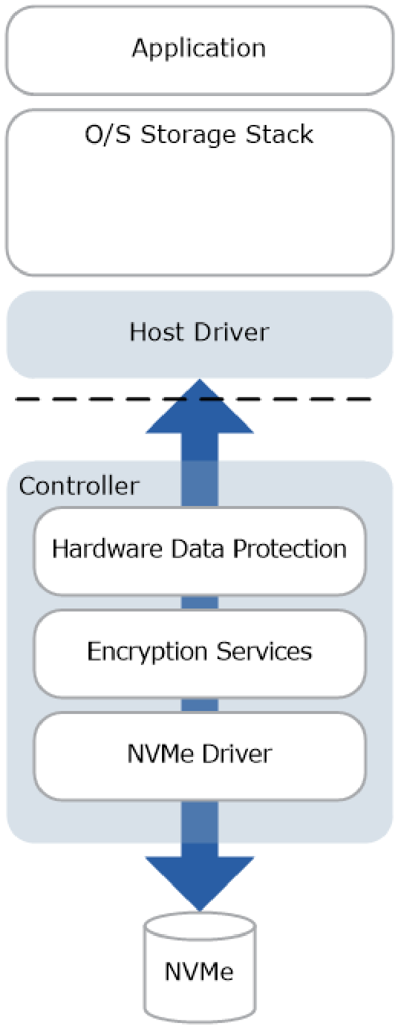
Software RAID Architecture



Option 2: Traditional Hardware RAID

Traditional hardware RAID relieves the parity management burden from the host. The use of traditional hardware RAID funnels all data to the RAID controller hardware prior to placing it on the drive's submission queue. Directing all data through the RAID controller adds complexity to the data path even when it is not necessarily needed.

Traditional Hardware RAID Architecture

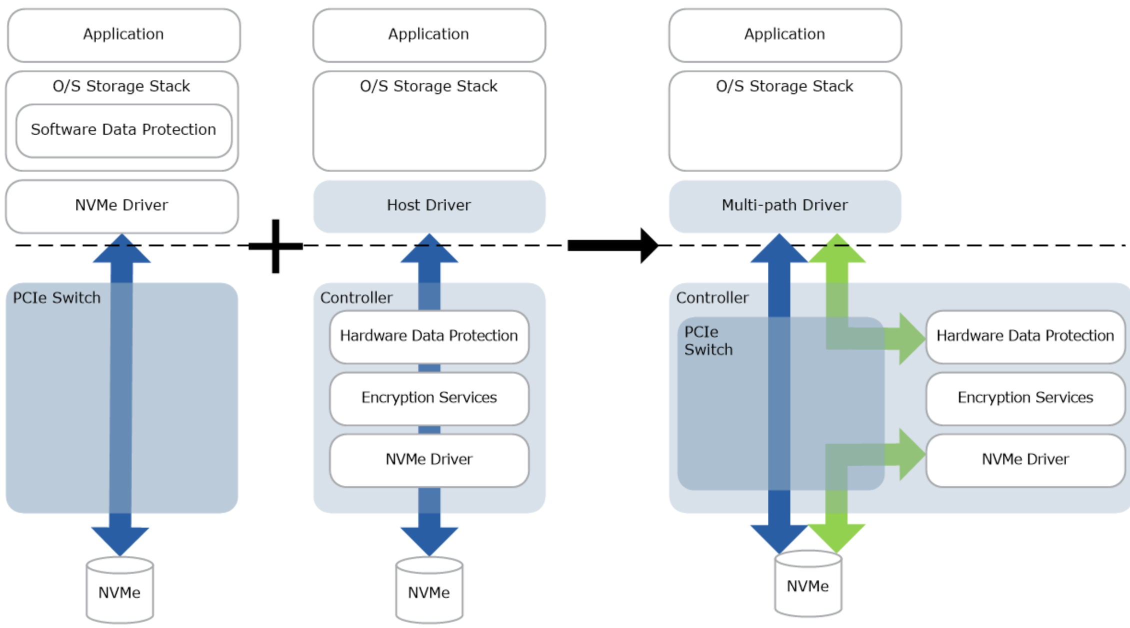


Traditional hardware RAID controllers offload host compute and memory resource consumption from the host but the controller is always present in the path to the drive, even in situations where offloads are not required. For example, writes that require parity generation need to be routed through the controller to take advantage of the offload but a read to the same logical volume would be lower latency if it could be issued from the host straight to the drive.

Option 3: NVMe-optimized Hardware RAID

Combining a multi-path driver with an embedded switch within the controller unlocks the best of previous architectures for NVMe drives. The embedded switch provides a streamlined data path through the adapter, unencumbered by firmware while maintaining the availability of hardware-accelerated advanced data services when the multi-path driver determines it is required. The multi-path driver intelligently manages data based on the data service requirements through either the switch or the RAID controller with negligible overhead.

Multi-path RAID Architecture

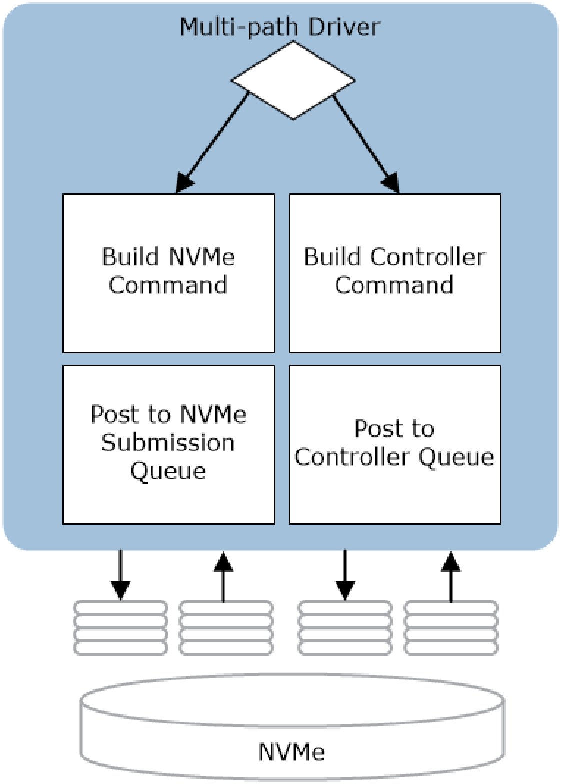


Multi-path Driver

The multi-path driver opportunistically submits commands directly to the drive submission queues via the embedded switch. When rich data services are required, such as parity or controller encryption, the multi-path driver forwards data to the value-add engines of the hardware controller.
NVMe direct path operations apply to all RAID modes for single column reads to all RAID volumes or single column writes to non-parity volumes when the array is in a good state. Path selection for commands is a nearly atomic operation and practically immeasurable in the I/O path. The subsequent command construction operates just as efficiently as a native NVMe driver in terms of instructions per cycle.

The following figure shows multi-path driver operation.

Multi-path Driver Operation



To determine which path to use, the driver performs a few lightweight checks such as:

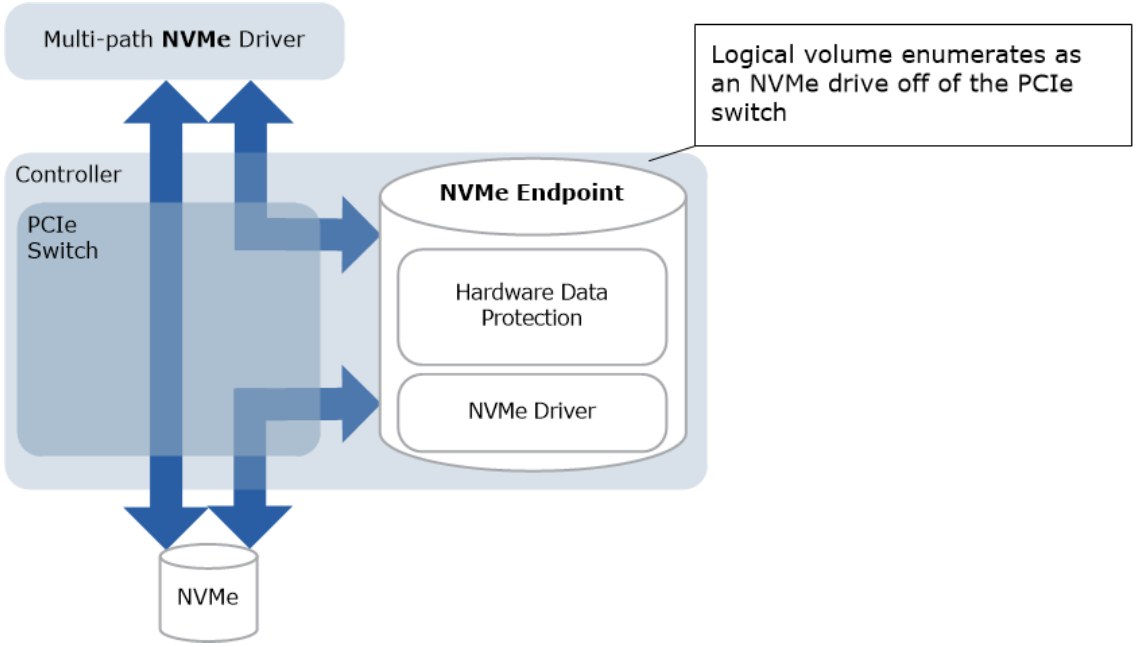
- Is the array in a good state?
- Is the command a single column I/O?
- Is the command a read or non-parity write?
- Is the data flow plaintext?

If all states are positive, the driver issues the NVMe command directly to the NVMe drive, through the embedded PCIe switch on the controller.

The Path to Standardized Drivers

In the prior section, the multi-path driver talks to the controller via a proprietary protocol and presents the controller-attached NVMe devices to the host operating system as logical SCSI volumes. A more standardized approach is also possible. The full offload path to the logical volume could be represented as a logical NVMe drive attached to the controller's PCIe switch, allowing a standard NVMe driver to enumerate and access the NVMe volume. A multi-path aware NVMe driver could then be developed to leverage the additional lower latency path directly to the drives via the controller's PCIe switch.

Standardized NVMe Driver Implementation



Conclusion

Expanded use of NVMe-based devices beyond high performance limited applications driven by the maturity of the NVMe-based storage ecosystem coupled with lowering economics, renews demand for data protection for storage. Demand for offloaded hardware RAID acceleration and rich data services have not dissipated but do require alteration from the traditional data path management to new optimized paths to maintain relevance in the NVMe market. Microsemi-based architectures deliver on the promise of optimized data paths for protected storage utilizing the latest in embedded switch technology and innovative driver support.

Microsemi Headquarters
One Enterprise, Aliso Viejo,
CA 92656 USA
Within the USA: +1 (800) 713-4113
Outside the USA: +1 (949) 380-6100
Sales: +1 (949) 380-6136
Fax: +1 (949) 215-4996
Email: sales.support@microsemi.com
www.microsemi.com

ESC-2181406