# An Introduction to NVMe

**Technology Paper**

Authored by:
Hermann Strass

NVM Express (NVMe) is a protocol for the transport of data over different media and for optimized storage in NAND flash. Peripheral Component Interconnect Express (PCIe) is currently the most used transport medium. Other media, like NVMe over Fabrics, are currently being standardized. NVMe is optimized for NAND flash chips. The protocol provides a high-bandwidth and low-latency framework to the storage protocol, but with flash-specific improvements.

## NVMe Protocol

NVMe is a scalable protocol optimized for efficient data transport over PCIe for storage on NAND flash, primarily deployed on PCIe solid-state drives today. It uses a simple streamlined, minimum set of only 13 commands, as listed in Table 1.

To optimize storage and retrieval, NVMe uses up to 64K commands per queue on up to 64K I/O queues for parallel operation. It has a paired submission-and-completion queue mechanism in host memory. Host software places commands into the submission queue. The NVMe controller places command completions into an associated completion queue. Multiple submission queues may report completions on a single completion queue, provided the controller supports arbitration with different priorities. Message-Signaled Interrupts Extended (MSI-X) and interrupt steering is supported as well. Optionally, support for many enterprise capabilities like end-to-end data protection (compatible with T10 DIF and DIX standards), enhanced error reporting, autonomous power state transitions for clients and hinting are included.

**Table 1. Simple Command Set – Optimized for NVMe (Source: NVM Express Org)**

| Only 10 admin commands are required | Only 3 I/O commands are required |
|---|---|
| ADMIN COMMANDS | NVM I/O COMMANDS |
| **Create I/O Submission Queue** | **Read** |
| **Delete I/O Submission Queue** | **Write** |
| **Create I/O Completion Queue** | **Flush** |
| **Delete I/O Completion Queue** | *Write Uncorrectable (Optional)* |
| **Get Log Page** | *Compare (Optional)* |
| **Identify** | *Dataset Management (Optional)* |
| **Abort** | *Write Zeros (Optional)* |
| **Set Features** | *Reservation Register (Optional)* |
| **Get Features** | *Reservation Report (Optional)* |
| **Asynchornous Event Requests** | *Reservation Acquire (Optional)* |
| *Firmware Activate (Optional)* | *Reservation Release (Optional)* |
| *Firmware Image Download (Optional)* | |
| *Format NVM (Optional)* | |
| *Security Send (Optional)* | |
| *Security Receive (Optional)* | |

The NVMe protocol progressed from version 1.0 to 1.2 in just four years. Version 1.0 featured end-to-end protection, queuing and security. It was enhanced with autonomous power transition, multipath I/O including Reservations and namespace sharing in version 1.1. Atomicity enhancements, live firmware updates, Namespace management, Host and Controller memory buffer, temperature thresholds and pass-through support were introduced in version 1.2. NVM subsystem statistics, sanitize command, Streaming and Attribute Pools will be part of NVMe version 1.3.

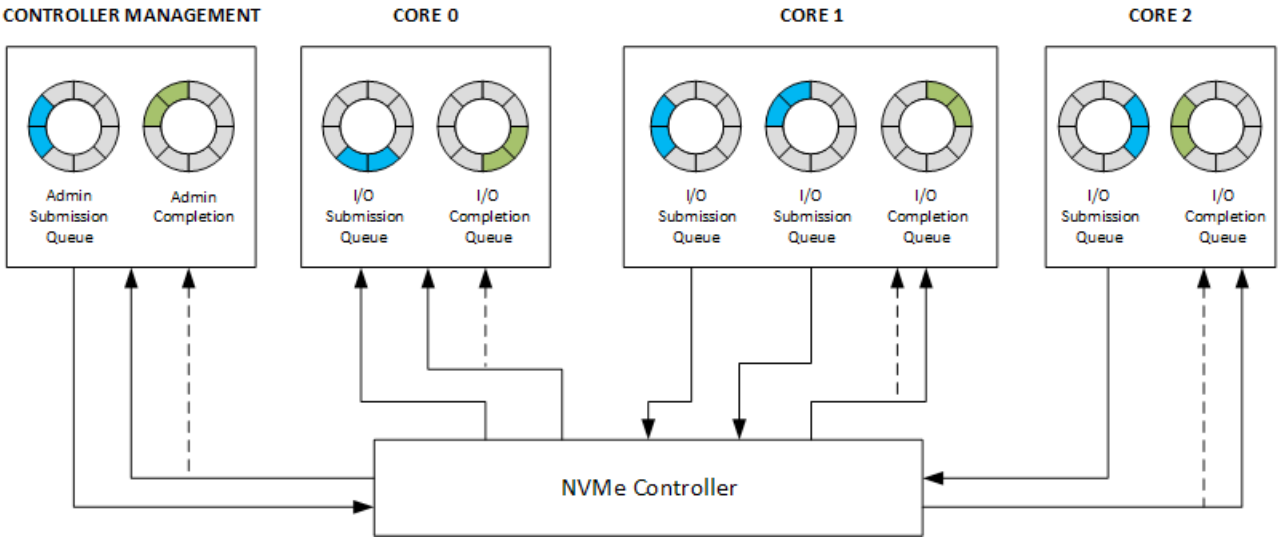


Figure 1. NVM Queue Architecture (Source: NVM Express Org)

The NVMe Management Interface (NVMe-MI) defines an out-of-band management that is independent of physical transport and protocol. It standardizes the out-of-band management to discover and configure NVMe devices and it maps the management interface to one or more out-of-band physical interfaces, like I2C (SMBus) or PCIe VDM.

The NVMe-MI carries the commands that are needed for systems management. Systems management includes the following elements:

- Inventory
- Configuration
- Health Status Monitoring
- Change Management

Single Root I/O Virtualization (SR-IOV) can virtualize one PCIe device so it appears to be many devices. This eliminates I/O bottlenecks in virtualized server environments and improves data rates up to full PCIe bandwidth.

There is a programming interface providing out-of-band management of NVMe Field Replaceable Units (FRU). An FRU is a circuit board, part or assembly that can be quickly and easily removed from a computer or other piece of equipment. A user or a technician can replace an FRU without having to send the entire product or system to a repair facility.

Some major NVMe protocol options are:

- Up to 64K I/O queues, with up to 64K commands per queue
- Priority associated with each I/O queue with well-defined arbitration mechanism
- All information for a 4KB read request is in the 64B command itself
- Efficient small random I/O operation
- Efficient and streamlined command set
- MSI/MSI-X and interrupt aggregation
- No un-cacheable/MMIO register reads required
- A maximum of one MMIO register write is necessary
- Multiple namespaces
- Efficient I/O virtualization architectures like SR-IOV
- Multipath I/O, including reservations
- Robust error reporting and management capabilities
- End-to-end data protection (DIF/DIX)

## Variants of the PCIe Form Factors

### AIC

PCIe Add-in-Cards (AIC) enable maximum system compatibility in existing servers and provide a reliable transport medium for a higher-power envelope and options for height and length.

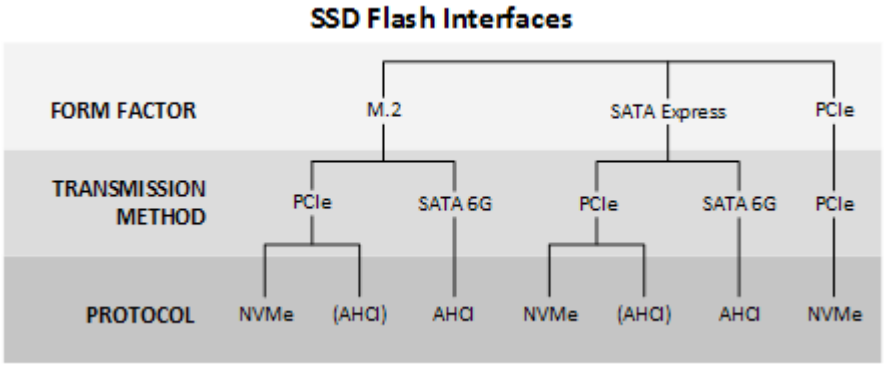


Figure 2. SSD Flash Interfaces and Protocols (Source: Elkomp & Techcon)

U.2

The U.2 connector and form factor (formerly SFF-8639 or Multifunction 6X Unshielded Connector) offers six lanes of high-speed data channels, supporting the following:

Single port SATA (one lane)

Dual port SATAe (two lanes)

Dual port SAS (two lanes)

MultiLink SAS (four lanes)

Four port PCIe (four lanes)

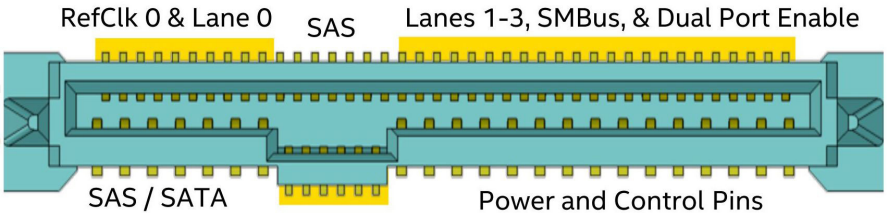


Figure 3. SFF-8639 (U.2) Connector (Source: SFF Committee)

It also accepts plugs in accordance with the SFF-8482 (dual SAS, EIA-966), SFF-8630 (MultiLink SAS) and SFF-8680 (high-speed dual SAS) standards. The SFF-8639 connector scheme has now been renamed as U.2 (Universal) in the style of M.2 (Mini).

Four lanes of PCIe or a mix of PCIe as well as two lanes of SAS and one lane of SATA are widely used in NVMe (U.2) drive bays to implement card cages of multiple flash boards in a 2.5-inch form factor (SFF-8223, EIA-720). 2.5-inch drive bay systems make up the majority of SSDs sold today because of ease of deployment, ability to hot-plug, serviceability and small form factor.

M.2

The M.2 (Mini) board form factor (plug-on modules on motherboards) is offered in many sizes, including widths of 12mm, 16.5mm, 22mm and 30mm and lengths of 16mm, 26mm, 30mm, 42mm, 60mm, 80mm and 110mm. Single-sided (S1 to S3) and double-sided (D1 to D3) versions with various component heights are defined. Other interfaces, like USB and $I^2C$, are also standardized. Not all combinations are used. Boards can also be keyed for SATA (socket 2) or PCIe ×4 (socket 3). Lengths of 42mm, 80mm and 110mm are most popular. M.2 boards provide the smallest footprint of PCIe in use for boot or for maximum storage density.

SATAe

The SATA Express (SATAe) connector supports drives in the 2.5-inch form factor, allowing for SSDs, hard drives or hybrid drives. It offers a combination of SATA and PCIe 3.0 channels. It is a cabled version of SATA compatible with SATA 3 (6Gb/s). The SATAe interface supports both PCIe and SATA storage devices by exposing multiple PCIe lanes and two SATA 3.0 (6Gb/s) ports through the same host-side SATAe connector. The PCIe lanes provide a pure PCIe connection with no additional layers of bus abstraction.

BGA

Some vendors have proposed a BGA solution (JEDEC MO-276) for small form-factor applications like two-in-one laptops.

## Inherent Advantages of NVMe

NVMe supports a large number of deep queues and commands per queue. It enables parallelism in multiple cores of the system while supporting the Non-Uniform Memory Architecture (NUMA). Lockless command submission and completion provides much better latency and bypasses all SCSI layers. Commands use simple PCIe register writes that allow much more flexibility in issuing 64-bit commands without any contention.

NVMe provides Controller Memory Buffer features that allow a host to prepare commands in controller memory. That means the controller no longer needs to fetch command buffers through PCIe reads. NVMe is a more streamlined interface, focused on passing memory blocks vs. SCSI commands. These features provide lower latency than other protocols. The NVMe arbitration mechanism improves flexibility in providing priority on a command basis for a better Service Level Agreement (SLA). The protocol enables reservations in data center applications. Major operating systems (OS) also include native NVMe driver support for ease of deployment.

NVMe also includes a host memory buffer that provides additional memory for its noncritical data structures. This allows DRAM-less NVMe controllers to use system memory. This is well suited for client and mobile NVMe controllers.

Lower I/O latency has been achieved by reducing block I/O stack overhead in the operating system, I/O process/resource locks and resource contentions. A version of a planned hot-swap function is available for 2.5-inch form factor devices. Modern processors have implemented NVMe PCIe 3.0 directly on the CPU. There is no need for an HBA or an I/O controller. This eliminates latencies and error possibilities. Reduced power consumption in 2.5-inch and M.2 form factors is possible: SATA is limited to 600 MB/s at 0.6W while PCIe Gen3 ×4 offers up to 3000 MB/s at 0.55W, more than five times the performance for the same power consumption.

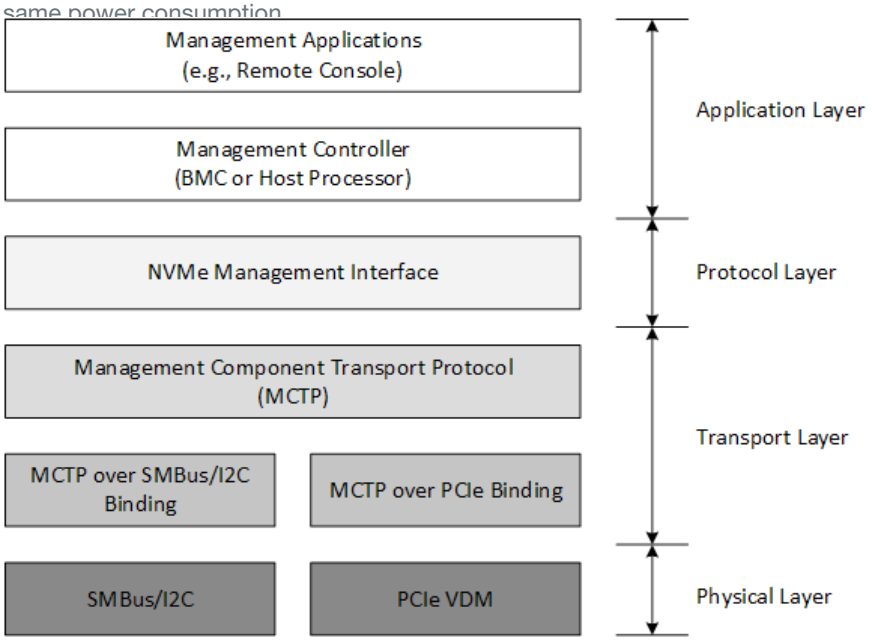


Figure 4. Four-Layer NVMe Management Hierarchy (Source: NVM Express Org)

## How NVMe Improves the User Experience

NVMe has many benefits compared to SATA or SCSI flash storage. Direct connection to the CPU provides lower latency compared to a connection via I/O controllers, multiplexers or storage networks. The scalable performance of up to 1GB/s per lane is achievable using industry standards like NVMe and PCIe 3.0. Currently, it supports up to 40 PCIe single-lane devices on a CPU system, but this is could increase to 48 in the future.

NVMe also provides end-to-end data protection via the Data Integrity Field/Data Integrity eXtension (DIF/DIX). Increased security has been achieved by using Enterprise, Opal and Opal light protocols from the TCG (Trusted Computing Group). Opal specifies minimum acceptable core capabilities of a storage device tailored for the PC client and value enterprise markets requirements. The Opal specifications provide a comprehensive architecture for putting storage devices under policy control as determined by the trusted platform host, the capabilities of the storage device to conform to the policies of the trusted platform and the life-cycle state of the storage device as a trusted peripheral. Opal protects the confidentiality of stored user data against unauthorized access once it leaves the owner's control by de-authentication following a power cycle.

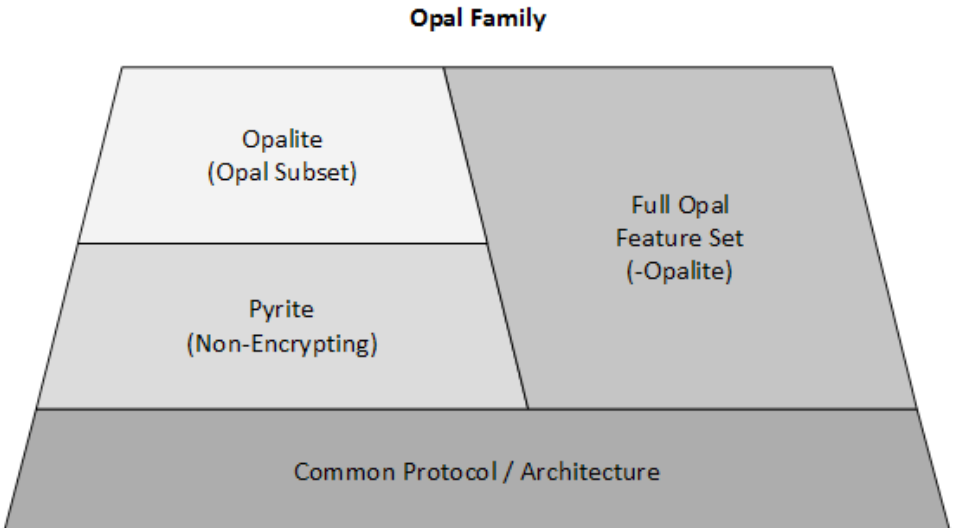


Figure 5: Opal Family Versions (Source: Trusted Computer Group)

The NVMe protocol supports different autonomous power state transitions like Low Power Link (L 1.2). The NVMe protocol is available on different form factors like M.2, SATAe, SFF-8639 (U.2) and SFF-8680, as outlined earlier.

## Valuable Features

### Hinting, Attribute Pools and Streams

The NVMe Dataset Management (DSM) commands provide an extensive hinting framework. DSM supports per-I/O (in-band) hinting as well as out-of-band hinting. Hinting includes I/O classification, policy assignment and policy enforcement. Objects and I/Os are identified or grouped by classifiers. Storage policies are assigned to classes of I/Os and objects, which are then enforced in the storage system. The upcoming NVMe standards include specifications for implementing Quality of Service (QoS) by defining bands, or Attribute Pools in NVMe parlance. Attribute Pools provide QoS by matching the type of storage (SSD, SAS/SATA HDD, SMR HDD or NVDRAM) to the application needs. This is covered in detail in a section below. Streams represent groups of data that belong to the same physical location and hence can be allocated and freed together. The upcoming NVMe standards for Streams specify Stream control commands and Stream identifiers for I/O write commands, which, in turn, serve as hints to identify the related group of data.

### Streams

NVMe lets applications specify logical blocks as part of separate streams.This information can be used by the device for physical media allocation techniques to minimize garbage collection. This might result in a reduced Write Amplification Factor by enabling the device to free all physical media associated with a stream together or other performance-related enhancements. If an application can group and store user data that has the same expected lifetime (for example, the host can erase all the logical blocks of a stream together, or unmap all of these LBAs at the same time), this eliminates unnecessary data processing and improves overall NAND life and performance. Coupled with this, the upcoming NVMe standards for Streams also specify Accelerated Background Operations that let the host schedule device maintenance at times that do not conflict with critical applications. If the host is able to predict idle time when there are few read and write requests to the controller, then the host can start or stop host-initiated accelerated background operations. As a result, accelerated background operations might be minimally overlapped with read requests and write requests from the host.

### Attribute Pool

An Attribute Pool is a band of media that has certain characteristics associated with capacity, performance, endurance and longevity. An NVMe device can support multiple Physical Stores to make use of different media to provide different QoS for some LBAs/namespaces. The device may use either caching or tiering methods to provide the desired QoS. To provide management for these multistore devices, the host specifies the requested QoS by using the Management Attribute Pools feature. Attribute Pools are defined per Namespace. A single Attribute Pool can span across multiple Physical Stores. Each namespace may have one or more Attribute Pools.

### Dual Port Drives

Shared Namespaces for dual-ported drives may exist in two domains, where each port is connected to a different server/domain. A multipath driver handles load balancing using a round robin approach, which allows for higher throughput. In this topology, the same host has two redundant paths to an SSD. In case of a path failure, the host can re-route all I/O commands to the surviving path.

Coordinated access to the drive is controlled through Reservations set via Persistent reserve/release (PR) commands. There are six types of reservation possible, which are listed in Table 2. At any given time, only one type of Reservation can be set:

**Table 2. Reservation Types**

| Reservation Type | Reservation Holder | | Registrant | | Non-Registrant | | Reservation Holder Definition |
|---|---|---|---|---|---|---|---|
| | Read | Write | Read | Write | Read | Write | |
| Write Exclusive | Y | Y | Y | N | Y | N | One reservation holder |
| Exclusive Access | Y | Y | N | N | N | N | One reservation holder |
| Write Exclusive – Registrants Only | Y | Y | Y | Y | Y | N | One reservation holder |
| Exclusive Access – Registrants Only | Y | Y | Y | Y | Y | N | One reservation holder |
| Write Exclusive – All Registrants | Y | Y | Y | Y | Y | N | All registrants are reservation holders |
| Exclusive Access – All Registrants | Y | Y | Y | Y | N | N | All registrants are reservation holders |

The reservations and registrations are persistent across all Controller Level Resets and NVM Subsystem Resets. A Reservation can only be cleared by preempt or preempt-and-abort commands. Log pages and AENs (Asynchronous Event Notifications) are supported for PR.

As an example, if a controller supports the Write Exclusive type of reservation, then only the host holding the reservation can write to NVM. However, both hosts can read NVM. This model acts like an active-passive type of reservation.

Typically, clusters prefer to use the Write Exclusive – Registrants only type of reservation, where registrants can also send write commands to NVM or shared namespace. This allows two hosts to send write commands concurrently. This model acts like an active-active type of reservation. In this mode, typically, both hosts coordinate to write to different locations in NVM (for example, different Logical Block Addresses).

### NVMe DIF/DIX support

NVMe end-to-end data protection is compatible with SCSI Protection Information, commonly known as the T10 DIF and SNIA DIX standards.

NVMe supports the same end-to-end protection types as DIF. The type of end-to-end data protection (Type 1, Type 2, or Type 3) is selected when a namespace is formatted and is reported in the Identify Namespace data structure.

Type1/Type3 DIX is supported at the namespace level. DIX Metadata is passed as a separate buffer. The NVMe controller only checks 2B of CRC for the Read and Write command. For Type1, the controller checks the last 4B of the LBA with the Expected Initial Logical Block Reference Tag (EILBRT) for read commands and Initial Logical Block Reference Tag (ILBRT) for write commands. When an NVMe controller detects a CRC mismatch, it reports an error to the host, which then handles the corrective action.

### Testing

Solid State System testing is in accordance with JEDEC JC-64.8 or SNIA SSS PTS and other specifications. The University of New Hampshire Interoperability Lab (UNH-IOL) has collaborated with the NVMe organization to deliver a robust interoperability program. The UNH-IOL has extensive experience with protocol and channel interoperability testing, as it has provided this type of service for many other systems with different protocols or channels.

## Summary

NVMe enables customers to utilize the full performance and latency potential of flash storage, like higher I/O performance, by efficiently supporting more processor cores, lanes per device, I/O threads and I/O queues. NVMe eliminates the SCSI and ATA I/O command overhead processing. NVMe implements simplified command processing because all the commands are the same size and in the same location or position. NVMe is less complex, more efficient, more serviceable and easier to use when compared to legacy systems.

seagate.com

AMERICAS Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, United States, 408-658-1000
ASIA/PACIFIC Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapore 569877, 65-6485-3888
EUROPE, MIDDLE EAST AND AFRICA Seagate Technology SAS 16–18, rue du Dôme, 92100 Boulogne-Billancourt, France, 33 1-4186 10 00