Study on NVMe

**Intro:**

NVM Express (NVMe) is a specification that defining how host software communicates with non-volatile memory across a PCI Express (PCIe) bus. It is the industry standard for PCIe Solid State Drives(SSDs) in all form factors(U.2, M.2, AIC, EDSFF). NVMe is the non profit consortium of tech industry leaders defining, managing and marketing NVMe technology. In addition to the NVMe base specification, the organizations hosts other specifications: NVMe over Fabrics (NVMe-oF) for using NVMe commands over network fabric and NVMe Management Interface (NVMe-MI) to manage NVMe/PCIe SSDs in servers and storage systems.

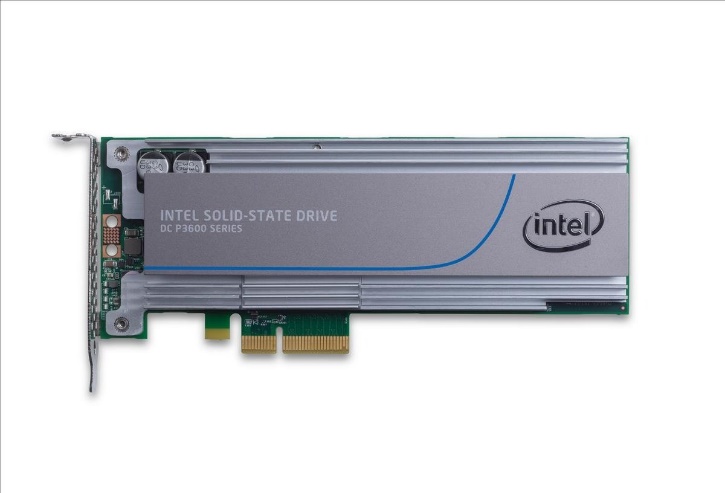
NVMe PCIe SSDs are able to achieve higher transfer speeds because of directly connected to the CPU using PCIe bus.

**SSD form factors:**

M.2: It is aka Next Generation Form Factor (NGFF) is a specification for internally mounted computer expansion cards and associated connectors. M.2 replaces the mSATA standard, which uses the PCIe mini card physical layout and connectors. M.2 is available in different widths and lengths makes it suitable for Ultrabook’s and tablets.



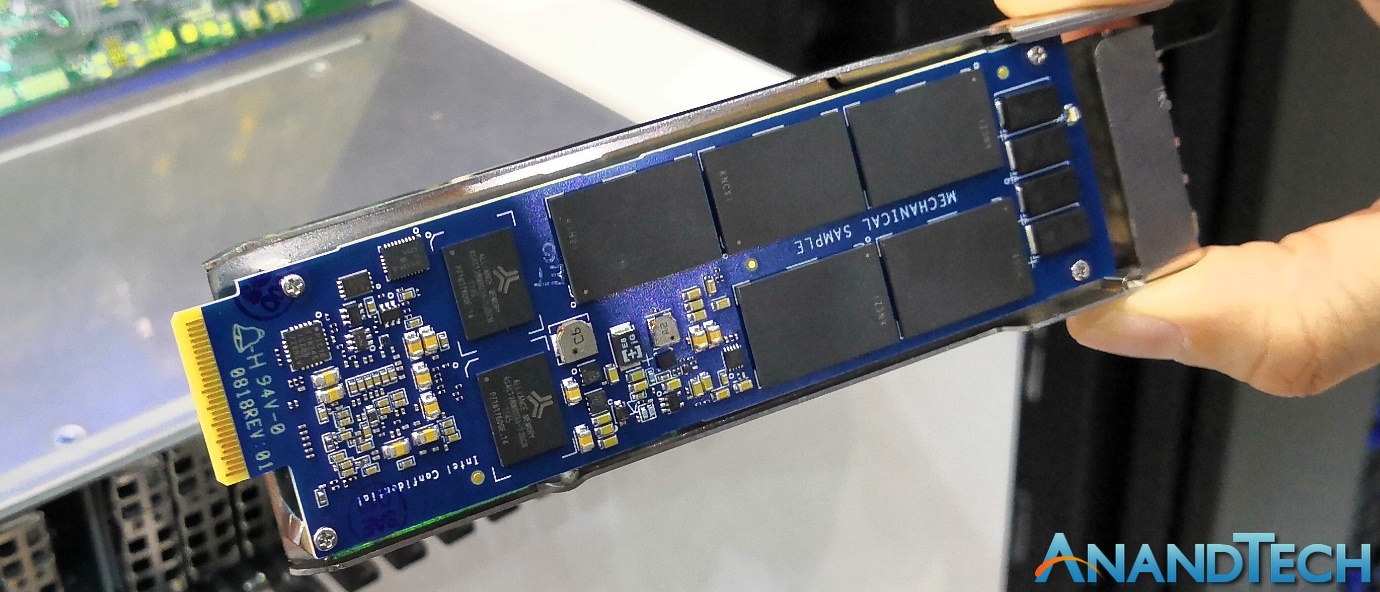
AIC: Add in card PCIe SSDs are similar to the graphics or audio add in cards that sit in PCIe slot.



U.2: It is formerly known as SFF-8639, is a computer interface standard for connecting SSDs to a computer. It covers the physical connectors, electrical characteristics, and communication protocols. It used four PCIe express lanes and two SATA lanes. It is designed to be used with enterprise market.



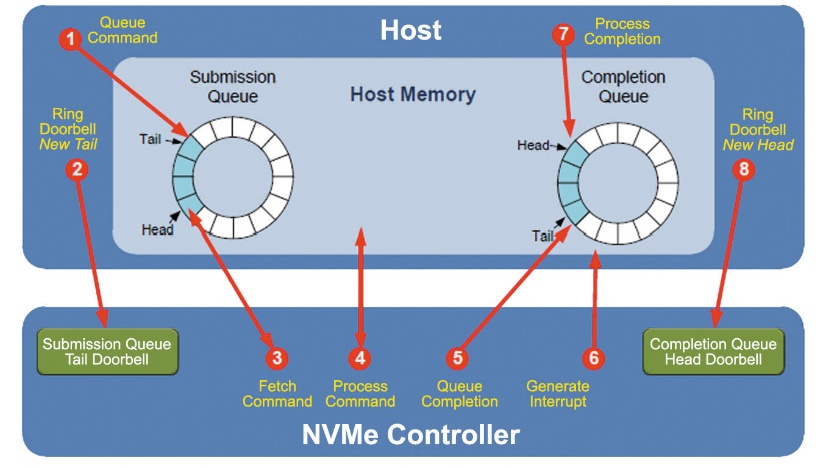
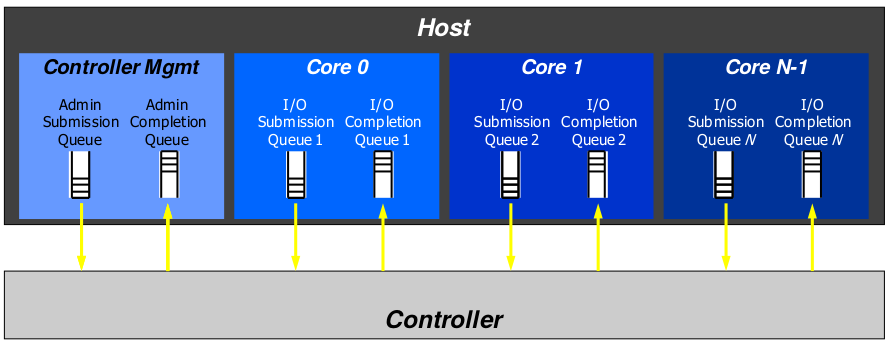
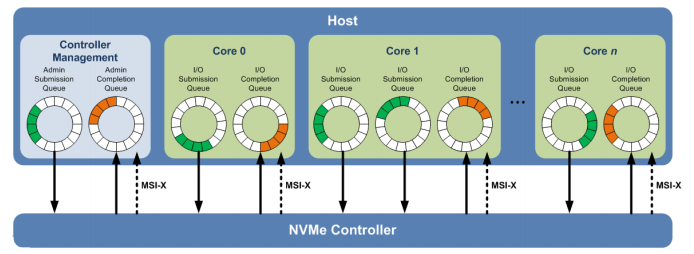
EDSFF: These SSDs are typically used in data centre solutions.



**Features:**

* Does not require uncacheable/MMIO register reads in the command submission or completion path
* Supports upto 65535 I/O queues, with each I/O queue supports up to 65535 outstanding commands.
* All information to complete a 4KiB read request is just included in 64B command itself, ensuring efficient small IO operations.
* Supports multiple namespaces.
* Robust error reporting and management capabilities.
* Supports multi path I/O and namespace sharing.

**Theory of Operation:**

* An NVMe controller is associated with single PCI function. The capabilities and settings that apply to this controller are indicated in the Controller Capabilities (CAP) register and the identify control data structure.
* A namespace is a quantity of non-volatile memory that may be formatted into logical blocks. NVMe controllers may support multiple namespaces that are referenced using a namespace ID. Namespaces are added and removed by using Namespace Management and Namespace Attachment commands.
* NVMe interface is based on pair of submission and completion queue mechanism. Commands are placed by host software in submission queue and completions are placed into associated completion queue by the controller. Multiple submission queues may use the same completions queue. All the submission and completion queues are allocated in memory.
* Admin commands are put into Admin submission queue.
* An I/O command set is used with an I/O queue pair. The specifications define one I/O command set named NVMe Command Set.
* Host software created queues, up to the maximum supported by the controller. Basically, the number of queues created is based on the system configuration and anticipated workload. Suppose on a quad core processor-based system, there will be one queue pair per core.
  + Queue Pair 1:1 -> single Submission Queue vs single Completion Queue
* Queue Pair n:1 -> multiple Submission Queues vs single Completion Queue
* Submission Queue (SQ) is a circular buffer with a fixed slot size that the host software uses to submit the commands for execution by the controller. The host software updates the appropriate SQ Tail doorbell register when there are multiple commands in the SQ. Then the controller fetches the commands from the SQ for execution. Each command in the queue is of size 64 bytes.
* Completion Queue (CQ) is a circular buffer with a fixed slot size that the host software uses for completions results of submitted commands. A completion command is uniquely identified by a combination of SQ identifier and command identifier that is assigned by host software.

**Multi-Path I/O and Namespace sharing:**

A multi path I/O refers to two or more completely independent paths between a single host and a namespace.

Namespace sharing refers to the ability for two or more hosts to access a shared namespace using different NVMe controllers.

Fig 1: NVMe controller with two namespaces

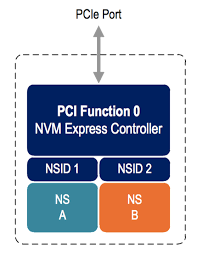


Fig 2: Two controllers with one port sharing namespaces

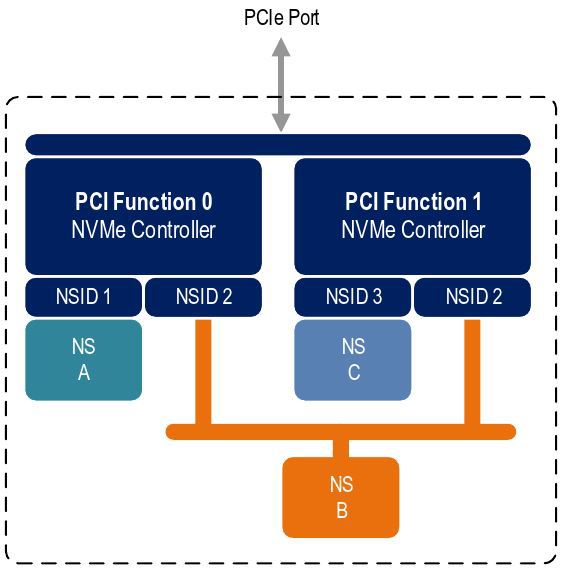


Fig 3: Two controllers with two ports sharing namespaces

