NVMe Hot-Plug Challenges and Industry Adoption

Tech Note by:
Austin Bolen

SUMMARY
Dell EMC understands that hot-plug operations for NVMe SSDs while the server is running are essential to reducing and preventing costly downtime.

PowerEdge 14G servers support a wide variety of hot-plug serviceability features, including:

Surprise insertion, which enables addition of NVMe SSDs to the server without taking the server offline.

Surprise removal on OSes that support it, which allows a user to quickly remove a faulty, damaged, or worn out NVMe SSD.

Dell EMC PowerEdge 14th-generation (14G) servers support a wide variety Reliability, Availability, Serviceability, and Manageability (RASM) features designed to enhance server uptime and reduce total cost of ownership, as shown in Figure 1 below:

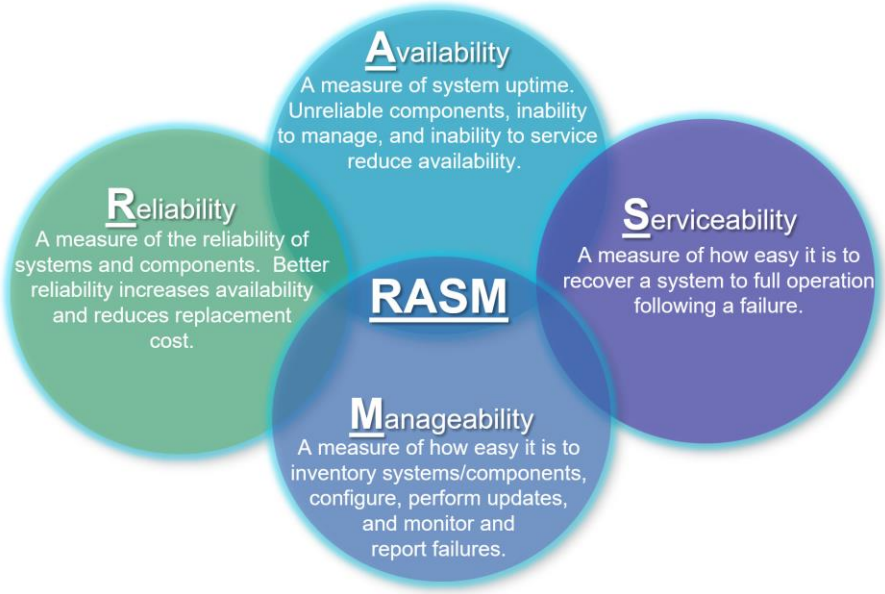

Figure 1 Reliability, Availability, Serviceability, and Manageability

One notable RASM feature supported on PowerEdge servers is the serviceability of Hard Disk Drives (HDDs) and Solid-State Drives (SSDs) and most recently NVM Express (NVMe) solid-state drives (SSDs). NVMe is an industry standard storage protocol designed to optimize performance of solid-state drives. Serviceability features allow NVMe SSDs to be added, removed, or replaced without the server having to be opened or turned off. This allows for easy replacement and/or re-provisioning.

NVMe SSDs in the U.2 2.5" form-factor are typically located in the front of PowerEdge servers which enables the easiest accessibility, however there are designs where these devices reside in the rear of the server. Refer to the Installation and Service Manual for your PowerEdge server for more details on the location and servicing of NVMe SSDs.

Serviceability is further enhanced by allowing U.2 2.5" NVMe SSD mounted in the front or rear of the server to be serviced while the server is powered on and running using an industry feature referred to as hot-plug which maximizes availability by minimizing costly server downtime. Hot-plug is broken down into two operations:

- **Hot Insert:** You insert an NVMe SSD into a running server.
  - **Surprise Insertion**: Prior to physically inserting the NVMe SSD, you do not notify the system that the NVMe SSD is about to be inserted.
- **Hot Removal:** You remove an NVMe SSD from a running server.
  - **Surprise Removal:** Prior to physically removing the NVMe SSD, you do not notify the system that the device is about to be removed.
- There are also orderly operations where operating system commands are used.
  - **Orderly Insertion**: Prior to physically inserting the NVMe SSD, you notify the system that the NVMe SSD is about to be inserted.
  - **Orderly Removal**: Prior to physically removing the NVMe SSD, you notify the system that the NVMe SSD is about to be removed.

PowerEdge servers and the operating systems supported on them support surprise insertion. There is no need to notify the system before hot-inserting an NVMe SSD.

Note: For surprise removal of any storage device (SAS, SATA, USB, NVMe, etc.), the user must ensure the data is not critical to the functioning of the system before removing the storage device. For example, a non-RAID boot storage device or swap file storage device could typically not be removed from a running system as doing so would likely crash the operating system.

Figure 2 Hot-Pluggable NVMe SSDs

The factors below also impact the ability to successfully hot-plug NVMe SSDs on PowerEdge servers:

- **Form Factor** – Hot-plug is only supported on U.2 2.5" form factor NVMe SSDs externally accessible in the front or rear of the server. PCIe Adapter Card NVMe devices do not support hot-plug.
- **Mechanicals** – PowerEdge servers are designed with high insertion count connectors on our backplane designs as well as the NVMe SSDs we use. For hot insertions, the NVMe SSD needs to be fully inserted. For hot removals, the NVMe SSD needs to be fully removed.
- **Number of NVMe SSDs supported** - All NVMe U.2 2.5" SSDs are hot pluggable, but only one at a time should be hot plugged.
- **Operational Times** – Hot-plug operations are only supported once the operating system has loaded. Hot plug operations are not supported when the OS is shutting down or in pre-boot.
- **Timing –** Hot-plug operations should be performed in a timely manner. Removal and insertion should be completed within 1 second.

We've discussed above what hot-plug is and why it is important to users. We will now go into details on the inter-dependencies of the operating system and BIOS to support hot-plug operations with NVMe SSDs.

For many storage device protocols, such as SAS, SATA, and USB, there is no need for orderly removal operations provided the data on the drives is not critical for continued operation of the system. For these protocols, surprise remove will suffice. Many operating systems, NVMe device drivers, and applications may not support surprise removal of NVMe SSDs.

Operating systems, drivers, and applications have many years of hardening to be able to reliably handle surprise removal of SAS, SATA, and USB storage devices. In all of these cases, there is a storage controller that acts as an intermediary between the storage device and the operating system, drivers, and applications. While the drives themselves are removed, the SAS, SATA, and USB storage controllers that the operating system, drivers, and applications talk to remain in place and are never removed. These controllers are shown above the hot-plug barrier in Figure 3.

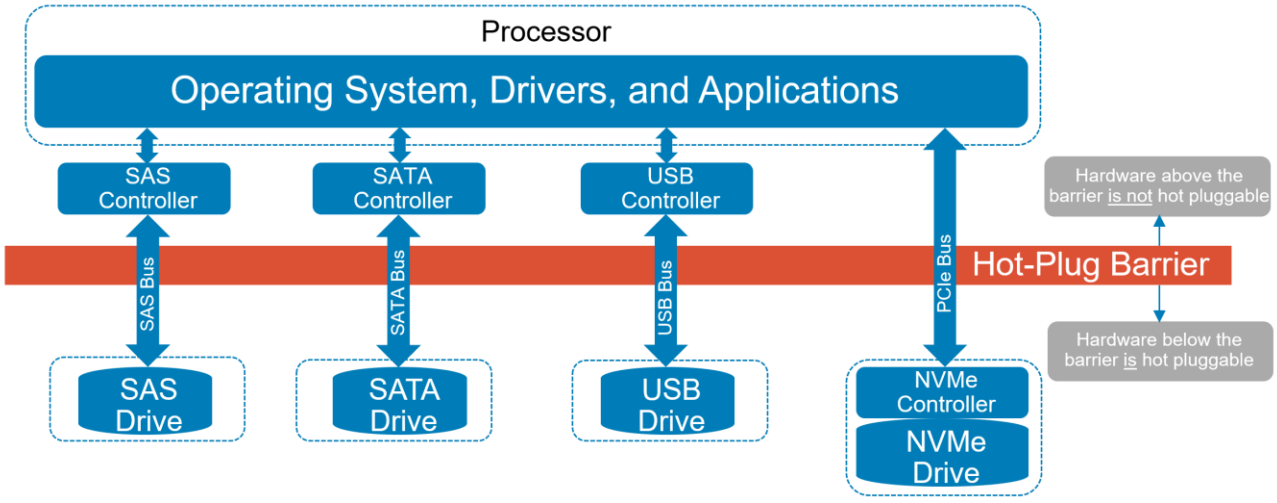

Figure 3 Storage Controller Hot-Plug Comparison

In NVMe, the storage controller was moved down on to the storage device below the hot-plug barrier as shown in Figure 3. An advantage of this approach is that it removes an added component layer when compared to the other storage solutions and helps NVMe to achieve such low latency accesses. However, this introduces a new model that operating systems, drivers, and applications had not dealt with before: the storage controller itself is removed when the storage device is removed.

Also note in Figure 3 that SAS, SATA, and USB have their own dedicated bus that have been architected for hot-plug. NVMe SSDs sit on the PCIe bus. The Conventional PCI bus architecture created in the 90s had no support for hot-plug. Afterwards, a hot-plug model referred to as the Standard Hot-Plug Controller (SHPC) model (https://members.pcisig.com/wg/PCI-SIG/document/download/8236) was added to Conventional PCI but required orderly removal and orderly insertion. When PCIe was introduced (the follow-on to the Conventional PCI/PCI-X busses) it adopted the SHPC orderly insert/remove model.

There was some rudimentary support for surprise removal added to PCIe but it was not architected with the complexities of NVMe SSDs in mind. Only recently has there been strong market demand for hot-plug PCIe devices. As a result, operating system vendors and application developers have not invested much effort into supporting this use case. As new protocols emerge that benefit from PCIe surprise removal like NVMe and Thunderbolt, operating systems and applications developers are starting to add more robust support, but it will take time for all operating systems and applications to complete the PCIe surprise remove support. As of the writing of this paper, the following outlines the state of support for NVMe SSD surprise removal in various PowerEdge server supported operating systems:

- Microsoft supports surprise removal of NVMe SSDs starting with Windows Server 2016.
- VMWare operating systems do not yet support surprise removal of NVMe SSDs. Full support for surprise removal of NVMe SSDs is expected in a future release of ESXi.
- The Linux server distributors like Red Hat Enterprise Linux, SUSE Linux Enterprise Server, and Ubuntu Server do not yet support surprise removal of NVMe SSDs. There are many developers actively submitting patches to the Linux kernel to harden the support for surprise removal of NVMe SSDs. Dell EMC continues to work with the open source Linux kernel community and the Linux server distributors and expect these operating systems to have full support for surprise removal of NVMe SSDs in future releases.

Many aspects of the system need to be modified in order to support surprise removal of NVMe SSDs. Dell EMC has made the changes at the server level (BIOS/UEFI System Firmware, iDRAC, backplanes, cables, etc.) and to Dell EMC applications/drivers (OpenManage Server Administrator, Dell Update Package, S140 Software RAID, etc.) to support surprise removal of NVMe SSDs. Dell EMC has also worked with the PCIe silicon vendors that provide PCIe root ports and PCIe switches used in PowerEdge servers to ensure they support surprise removal of NVMe SSDs.

Dell EMC qualified NVMe SSDs also support features needed for surprise removal such as power-loss protection (PLP) which ensures they can commit data in volatile memory buffers on the NVMe SSD to persistent memory on a power loss due to surprise removal or other conditions. When using NVMe SSDs not qualified by Dell EMC, the user should check with the vendor of those NVMe SSDs to ensure they support surprise removal.

For operating systems or applications that do not support surprise removal of NVMe SSDs, Dell EMC management tools such as OpenManage Server Administrator and iDRAC provide the user with an option to do an orderly removal via the “Prepare to Remove” task. Figure 4 on the following page shows the “Prepare to Remove” task for an NVMe SSD in OpenManage Server Administrator. For more details on the “Prepare to Remove” task refer to the User’s Guide for OpenManage Server Administrator and iDRAC (https://www.dell.com/support/home/us/en/04/products/software_int/software_ent_systems_mgmt) and Dell PowerEdge Express Flash NVMe PCIe SSD 2.5 inch Small Form Factor (http://topics-cdn.dell.com/pdf/dell-poweredge-exp-fsh-nvme-pcie-ssd_users-guide_en-us.pdf). These management tools will attempt to determine if the

NVMe SSD is in use and warn the user if so. They cannot detect all cases where an NVMe SSD is in use and so the user should verify the NVMe SSD is no longer in use prior to removing it. Some operating systems may prevent orderly removal of NVMe SSDs that are still in use.

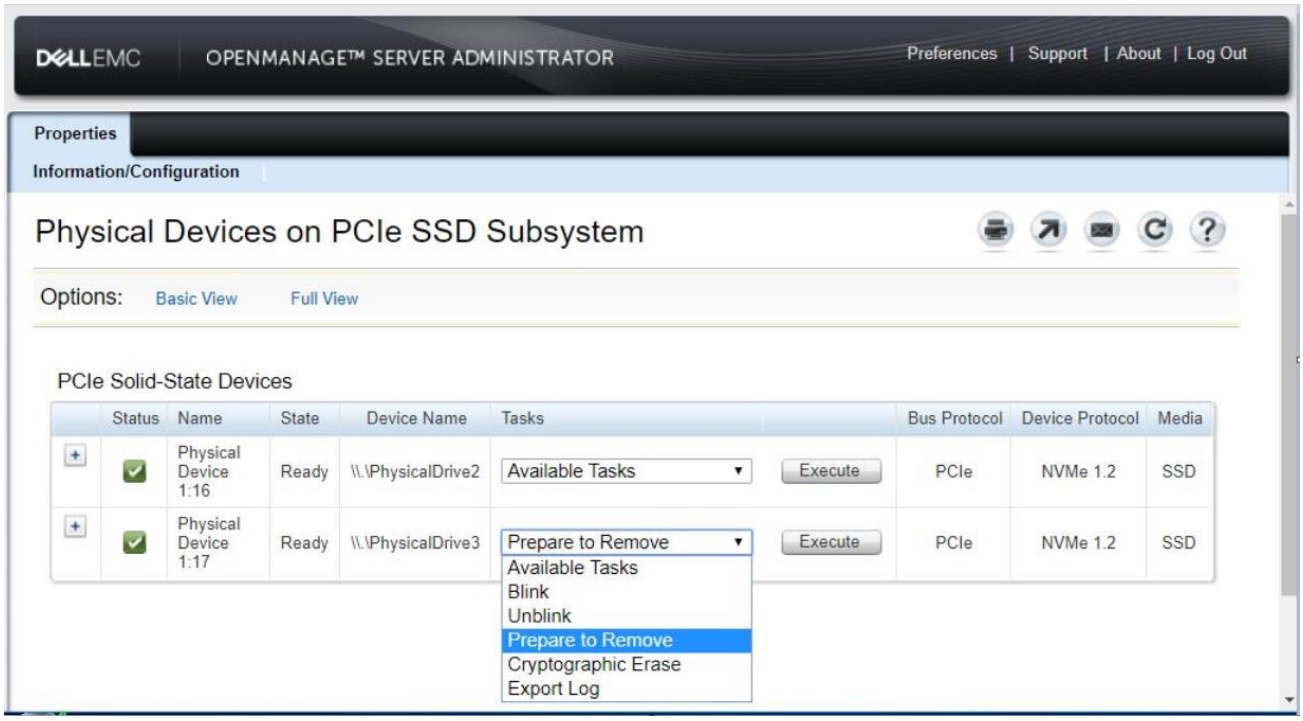

Figure 4 Prepare to Remove NVMe SSD

Users will need to check with the vendor of any operating system or third-party application that accesses NVMe SSDs to determine if it supports surprise removal of NVMe SSDs. For operating systems or third-party applications that do *not* support surprise removal of NVMe SSDs, users should perform an orderly removal as described above.

Dell EMC is also working with various industry standards bodies such as PCI-SIG (https://pcisig.com/) and the ACPI Specification Working Group (https://www.uefi.org/workinggroups), silicon providers, operating system vendors, and other OEMs to define new industry standard mechanisms to further improve support for NVMe hot-plug operations in the future.

Conclusions

Dell EMC PowerEdge 14th-generation (14G) servers support a wide variety of hot-plug serviceability features for NVM Express (NVMe) Solid-State Drive (SSDs) that address RASM and improve TCO. Surprise insertion is supported to allow adding NVMe SSDs to the server without taking the server offline. For operating systems that support it, surprise removal is supported to allow a user to quickly remove faulty, damaged, or worn out NVMe SSDs. Dell EMC understand that hot-plug operations for NVMe SSDs while the server is running reduces costly downtime and are driving the industry to improve user experience.