DELL Technologies

Technical White Paper

NVMe Surprise Removal on Dell EMC PowerEdge servers running Linux operating systems

Abstract

This white paper describes the support for Non-Volatile Memory Express (NVMe) Surprise Removal on Dell EMC PowerEdge servers running supported Enterprise Linux operating systems.

March 2021

Revisions

Date	Description
October 2020	Initial release
December 2020	Document updated with NVMe surprise removal information for Ubuntu LTS 20.04.01 Server
March 2021	Document updated with NVMe surprise removal information for Red Hat Enterprise Linux 8.2

Acknowledgements

Author: Narendra K

Support: Austin Bolen, Gurupreet Kaushik, Sherry Keller

Table of contents

Executive summary

NVMe devices are being used more widely, and features such as surprise removal are important to the continuous availability of the server and serviceability needs. Surprise removal allows you to remove a device from the server without prior notification. This white paper outlines the best practices that are to be followed for the surprise removal of NVMe devices running supported Linux operating systems on supported Dell EMC PowerEdge servers. Both supported and unsupported scenarios and known issues encountered while performing surprise removal on Linux operating systems are documented in this white paper.

1 Introduction

As NVMe devices are being used more widely, they must provide enterprise functionality such as surprise removal that you rely on. Surprise removal enhances the serviceability of NVMe devices by eliminating additional steps required to prepare the devices for orderly removal and ensures availability of servers by eliminating server downtime.

1.1 Audience and scope

The intended audience for this white paper includes IT administrators and those using hot-pluggable NVMe devices on Dell EMC PowerEdge servers running supported enterprise Linux operating systems.

1.2 Terminology

Hot insertion: Connecting the NVMe device to the server when the Linux operating system is booted up.

Surprise removal: Removing the NVMe device from the Linux operating system without notifying the operating system beforehand.

Orderly removal: Removing the NVMe device from the server after completing the prerequisites, such as suspending all processes accessing the NVMe device and quiescing all I/O operations accessing the NVMe device.

Hot swap: Replacing an existing NVMe device with a new NVMe device from the same or different vendor while the host operating system is booted. Hot swap is a surprise removal or orderly removal followed by a hot insertion operation with a different NVMe device.

1.3 Command-line utilities used for verifying surprise removal of NVMe devices

The following command-line utilities that are available in the enterprise Linux operating systems are used to verify hot-plug operations:

- nvme-cli
- lspci
- lsblk

2 Surprise removal of NVMe devices

2.1 Supported and unsupported scenarios for surprise removal of NVMe devices

The following table describes the supported and unsupported scenarios while performing surprise removal of NVMe devices.

Table 1 Supported and unsupported scenarios for surprise removal of NVMe devices

Supported scenarios	Unsupported scenarios
Surprise removal of a single NVMe device at a time is supported. The following requirements ensure successful surprise removal of NVMe devices: • Surprise removal must be performed within one-second period, as a slower surprise removal may cause the operating system to crash. • To avoid an operating system crash, a fifteen-second time interval should be provided between successive hot-plug operations to ensure that the operating system, applications, and drivers have enough time to fully handle the operation.	• Performing surprise removal of the drive that has the operating system installed or the drive that has a swap partition. • Performing surprise removal when the operating system is booting up. • Performing surprise removal of an NVMe device when another NVMe device is being hot inserted, or within 15 seconds of another NVMe device being hot inserted. • Performing surprise removal of two or more NVMe devices serially without a fifteen second time interval between the surprise removals. • Surprise removal of an NVMe device that is either directly or partially assigned to a virtual machine.

Note: Specific solutions may have additional requirements to perform successful surprise removal. For more information, see your solution documentation.

2.2 Identifying the NVMe device slot and verifying surprise removal

This section describes a scenario where /dev/nvme0n1 is the device to be surprise removed. The slot numbers used in this section are specific only to this use case.

Note: Surprise removing an NVMe device that is in use may result in data loss. It is recommended that you create a data backup before surprise removing the NVMe device.

To perform surprise removal of an NVMe device:

1. Use the command `nvme list` to list the NVMe devices connected to the server.
2. Use the command `nvme list-subsys` to retrieve the PCI bus/device/function number of the /dev/nvme0n1 device.
3. Determine the PCIe slot number using the PCI bus/device/function number and surprise remove the NVMe device from slot 22.

```
localhost:~ # cat /sys/bus/pci/devices/0000\:3d\:00.0/label
PCIe SSD in Slot 22 Bay 1
```

Figure 1 Determining the PCIe slot number of the /dev/nvme0n1

4. To verify that the operating system successfully unregisters the device:

   a. Use the command `nvme list` to list the connected devices and verify that the /dev/nvme0n1 is not listed.
   b. Use the command `lspci` to verify PCIe device 0000:3d:00.0 is not listed.
   c. Use the command `lsblk` to verify that the /dev/nvme0n1is not listed.

⚠ | CAUTION: The operating system might crash if subsequent hot-plug operations are not performed at time intervals of at least fifteen seconds.

2.3 Platform and operating system support summary

The following table lists the Dell EMC PowerEdge servers and the Linux operating systems that support NVMe surprise removal.

Table 2 Supported Dell EMC PowerEdge servers and Linux operating systems that support NVMe surprise removal

Dell EMC PowerEdge generation	SUSE Linux Enterprise Server Service Pack 2		Red Hat Enterprise Linux 8.2*		Ubuntu LTS 20.04.01	
	Supported	**Unsupported**	**Supported**	**Unsupported**	**Supported**	**Unsupported**
Intel Skylake and Cascade Lake SP CPU based yx4x servers	• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal	
AMD Naples CPU based yx4x servers	• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal	
AMD Rome CPU based yx5x servers	• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal		• Hot insertion • Orderly removal • Surprise removal	

Note: Linux upstream kernel version 5.7 and later have hot-plug related patches that enhance hot-plug user experience.

***Note:** The minimum kernel version required for surprise removal is version kernel-4.18.0-193.13.2.el8_2.x86_64.

3 Known issues with NVMe surprise removal

The following section describes the known issues encountered when surprise removal is performed on servers running supported Linux operating systems.

3.1 SUSE Linux Enterprise Server Service Pack 2

3.1.1 MD RAID layer is not notified of the surprise removal of Samsung NVMe devices

Description: When a virtual disk is created on the MD RAID layer using Samsung NVMe device, the MD RAID layer is not notified of the surprise removal of the NVMe drive. The output of the `mdadm -D` command displays an incorrect status of the MD RAID virtual disk. The issue is observed on Dell Express Flash PM1725a, PM1725b, Enterprise NVMe agnostic devices. Only the array status reporting is incorrect, however when I/O operations are performed, I/O errors are observed as expected and the filesystem changes to read-only.

Cause: The issue is observed while handling devices which showcase multipath capability.

Workaround: Pass the multipath=N module parameter to the nvme_core driver.

3.1.2 Status of the RAID 0 logical volume is displayed as Available when one of the members of the RAID array is surprise removed

Description: When Logical Volume Manager (LVM) is used to create a RAID 0 array and a member of the RAID array is surprise removed, the `lvdisplay` command shows the logical volume (LV) status as 'Available'.

Solution: Use the command `lvs -o +lv_health_status` to check the status of the RAID array. The command displays the output `Partial` when a member of the RAID array is removed. For more information, see SUSE Linux Enterprise Server Knowledge Base article 19716.

3.1.3 LVM does not activate a free physical volume when one of the NVMe devices is surprise removed

Description: When one of the members of a RAID 1 LVM array is surprise removed, the LVM does not replace the removed device with a free physical volume (PV) that is available in the volume group.

Cause: The issue is related to the handling of failover logic in the LVM.

Workaround: The command `lvconvert --repair` can be used to add the free PV to the RAID 1 LVM array.

Solution: The issue is resolved in the following Program Temporary Fix: www.ptf.suse.com/sle-modulebasesystem-15-sp2/20119/x86_64/20200820.

3.1.4 /proc/mdstat and mdadm -D commands display incorrect statuses when two NVMe devices are surprise removed from a RAID 5 MD array

Description: When two of three NVMe devices are surprise removed from a RAID 5 MD array, the command `cat/proc/mdstat` displays the array status incorrectly as `active`. Similarly, when the status of the MD RAID is queried using the `mdadm -D /dev/mdN` command, the number of `active` and `working` devices displayed is two. Only the status of the array reported is incorrect however, when I/O operations are performed, I/O errors are observed as expected.

Cause: When the number of devices that are surprise removed exceeds the number of devices that are required for the array to function, the MD status is not updated.

3.2 Red Hat Enterprise Linux 8.2

3.2.1 Dmesg displays error messages when NVMe device is surprise removed

Description: Dmesg or /var/log/messages show the following error messages after an NVMe device is unbound from the NVMe driver and surprise removed:

```
kernel: pcieport 0000:b0:06.0: Timeout waiting for Presence Detect
kernel: pcieport 0000:b0:06.0: link training error: status 0x8001
kernel: pcieport 0000:b0:06.0: Failed to check link status
```

The issue is a cosmetic issue and can be ignored.

Applies to: Red Hat Enterprise Linux 8.2 and later

Cause: The error that is displayed is due to an issue with the pciehp driver.

3.2.2 Status of the RAID 0 logical volume is displayed as Available when one of the members of the RAID array is surprise removed

Description: When Logical Volume Manager (LVM) is used to create a RAID 0 array and a member of the RAID array is surprise removed, the `lvdisplay` command shows the logical volume (LV) status as 'Available'.

Solution: Use the command `lvs -o +lv_health_status` to check the status of the RAID array. The command displays the output `Partial` when a member of the RAID array is removed.

3.2.3 /proc/mdstat and mdadm -D commands display incorrect statuses when two NVMe devices are surprise removed from a RAID 5 MD array

Description: When two of three NVMe devices are surprise removed from a RAID 5 MD array, the command `cat/proc/mdstat` displays the array status incorrectly as `active`. Similarly, when the status of the MD RAID is queried using the `mdadm -D /dev/mdN` command, the number of `active` and `working` devices displayed is two. Only the status of the array reported is incorrect however, when I/O operations are performed, I/O errors are observed as expected.

Cause: When the number of devices that are surprise removed exceeds the number of devices that are required for the array to function, the MD status is not updated.

3.3 Ubuntu LTS 20.04.01

3.3.1 The name of the NVMe device may change when it is hot inserted after a surprise removal

Description: If an NVMe device is hot inserted after it was previously surprise removed when I/O operations are accessing the device, the name of the NVMe device may change or will not retain the same name that is assigned prior to surprise removal. Dmesg displays the following messages:

```
kernel:  nvme nvme3: failed to mark controller CONNECTING
kernel:  nvme nvme3: Removing after probe failure status: -16
```

The functionality of the NVMe device is not affected.

3.3.2 NVMe devices are enumerated in namespace 2 when hot-inserted into the server after being surprise removed

Description: When an NVMe device from a RAID 1 MD array is hot inserted after being surprise removed, the device is enumerated in namespace 2 although only one namespace is enabled. The device is named as nvme2n2 instead of nvme2n1. This issue is observed on Dell Express Flash PM1725a device. The functionality of the NVMe device is not affected.

Workaround: Pass the multipath=N module parameter to the nvme_core driver.

3.3.3 Status of the RAID 0 logical volume is displayed as Available when one of the members of the RAID array is surprise removed

Description: When Logical Volume Manager (LVM) is used to create a RAID 0 array and a member of the RAID array is surprise removed, the `lvdisplay` command shows the logical volume (LV) status as 'Available'.

Solution: Use the command `lvs -o +lv_health_status` to check the status of the RAID array. The command displays the output `Partial` when a member of the RAID array is removed.

3.3.4 /proc/mdstat and mdadm -D commands display incorrect statuses when two NVMe devices are surprise removed from a RAID 5 MD array

Description: When two of three NVMe devices are surprise removed from a RAID 5 MD array, the command `cat/proc/mdstat` displays the array status incorrectly as `active`. Similarly, when the status of the MD RAID is queried using the `mdadm -D /dev/mdN` command, the number of `active` and `working` devices displayed is two. Only the status of the array reported is incorrect however, when I/O operations are performed, I/O errors are observed as expected.

Cause: When the number of devices that are surprise removed exceeds the number of devices that are required for the array to function, the MD status is not updated.

4 Summary

This white paper describes the concept of NVMe surprise removal and provides guidance on how to perform surprise removal on supported enterprise Linux operating systems on supported Dell EMC PowerEdge servers. The step-by-step instructions for performing NVMe surprise removal are documented with guidelines to be followed for successful surprise removal of NVMe devices. This document will be updated if there is a change in the support offered for surprise removal or if there are any major enhancements to the scenarios involving this feature. Further known issues related to surprise removal will be updated on the respective release notes document published on the operating system documentation page of www.dell.com/support.

5 References

- Dell Express Flash NVMe PCIe SSD User's Guide
- SUSE Linux Enterprise Server Certification Matrix for Dell EMC PowerEdge Servers
- Dell EMC PowerEdge Systems Running SUSE Linux Enterprise Server 15 Release Notes
- Ubuntu Server 20.04 LTS for Dell EMC PowerEdge Servers Release Notes
- RedHat Enterprise Linux Certification Matrix
- Dell EMC PowerEdge Systems Running Red Hat Enterprise Linux 8 Release Notes