Lenovo

Using Intel VMD and Intel VROC NVMe RAID on Lenovo ThinkSystem Servers

Introduces the architecture of Intel VMD and VROC technology

Describes the support of Intel VROC NVMe RAID on ThinkSystem servers

Explains how to enable the VROC Premium license for non-Intel NVMe drives

Provides instructions on how to configure UEFI to create a VROC RAID volume in UEFI using NVMe drives

Kelvin Shieh

LENOVO PRESS

Abstract

Intel Virtual RAID on CPU (VROC) is a RAID offering implemented using Intel Volume Management Device (VMD), a hardware architecture available in Intel Xeon Scalable Processors. Intel VROC is an enterprise solution designed to offer RAID functionality to NVMe SSDs and SATA SSDs connected to the processor without needing any additional RAID adapter or HBA.

Intel VROC provides robust Hot Plug support and LED status light management, and brings enterprise reliability, availability, and serviceability (RAS) features to NVMe SSDs, enabling you to deploy next-generation storage with confidence.

This document provides a brief introduction to the Intel VMD / VROC architecture and technology from hardware and software perspectives, and compares VROC to other conventional storage technologies. The paper also include detailed instructions on how to set up VMD on a Lenovo® ThinkSystem™ server, both in UEFI and in Red Hat Enterprise Linux 8.2 (RHEL 8.2). Finally, the paper discusses the known limitation of using VMD / VROC with prior Lenovo ThinkSystem servers.

This paper is intended for IT administrators. Readers are expected to have the basic knowledge of OS and storage deployment.

Contents

Introduction

Intel Virtual RAID on CPU (Intel VROC) is an enterprise RAID solution, specifically designed for NVMe SSDs connected directly to the CPU. Intel VROC is implemented using Intel Volume Management Device (Intel VMD), a hardware architecture implemented in Intel Xeon Scalable Processors. Intel VMD enhances the PCIe lanes supplies by the processor for dependable NVMe connections. Intel VROC uses Intel VMD for a simpler NVMe RAID solution that doesn't require an additional RAID adapter or HBA.

Intel VMD and Intel VROC drivers are delivered together as part of an enterprise storage stack. In Windows and Linux environments, these packages deliver LED Management and Hot-plug capabilities for direct attached RAID 0, 1, 5, and 10 arrays. If RAID is not needed, then Intel VROC can be used in pass-through mode to turn on Intel VMD Domains only. In VMware environments, Intel VMD can be used in both vSAN and direct attached storage modes, with the added benefit of RAID 1 for a redundant boot volume.

More details about the benefits of this Intel VMD / VROC technology can be found at:

https://www.intel.com/content/www/us/en/architecture-and-technology/intel-volume-management-device-overview.html

Architecture

In this section we compare the architecture of a traditional storage configuration and the Intel VROC configuration.

Conventional RAID storage adapter architecture

A conventional RAID adapter is an expansion card that is installed in a PCIe slot in the server and is connected via cables to storage devices such as hard disk drives (HDDs) and solid state drives (SSDs), either internally to the server, or external to storage enclosures.

RAID adapters support specific types of devices, such as Serial ATA (SATA) and Serial Attached SCSI (SAS). A RAID adapter combines multiple physical drives connected to it into virtual or logical drives. The operating system then recognizes each logical drive (in reality, at least two physical devices) as a single physical drive. Each logical drive that the RAID adapter creates has data protection built into it in the form of a RAID level.

The level of RAID assigned to an individual logical drive depends on the number of physical drives included (some RAID levels require minimum and/or maximum numbers of drives) and whether system performance or data security is the priority. A RAID adapter may also add additional cache memory to improve read and write performance.

Intel VMD / VROC architecture: Hardware view of the technology

Intel VMD is a new technology introduced with the Intel Xeon processor Scalable family primarily to improve the management of high-speed SSDs. SSDs were previously attached to a SATA or other interface types of adapters, which was managed through software. Intel VMD uses hardware to mitigate the management issues rather than completely relying on software.

The Intel Xeon Scalable family of processors (Gen 1, 2 and 3) has onboard capabilities that works to provide a quick access to the directly-attached NVMe SSDs on the PCIe lanes of the platform. The VMD device is embedded within the processor and thus no additional RAID adapter is required. These processors have a new hardware architecture, allowing NVMe SSDs to connect via PCIe connections and directly managed by processor.

The Intel VROC/VMD architecture is shown in Figure 1.

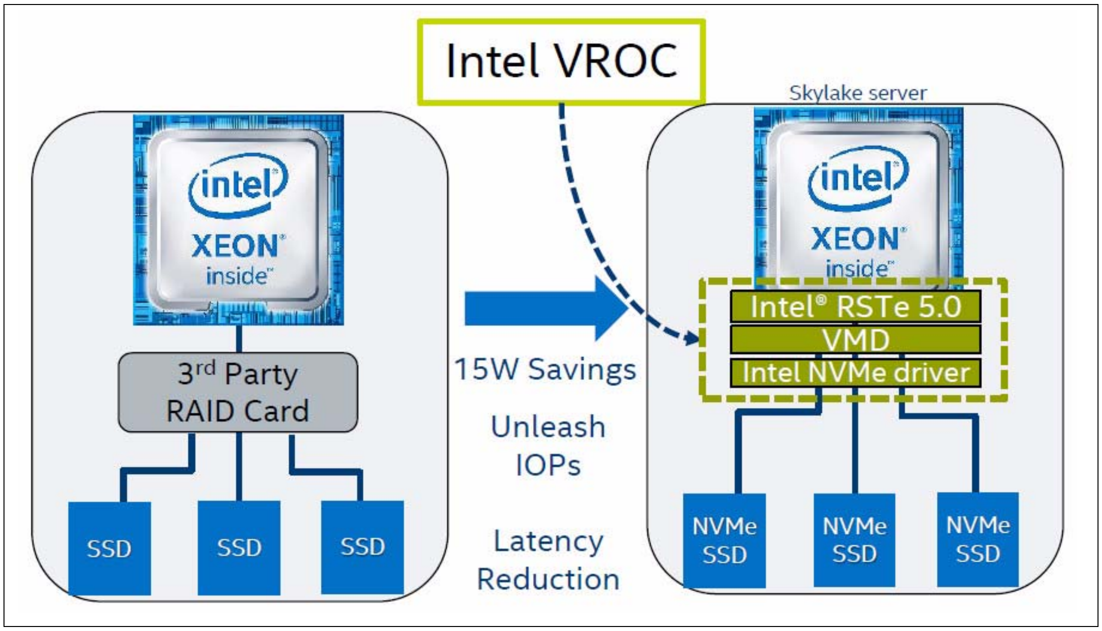


Figure 1 Comparing a traditional storage architecture to Intel VROC

For more details about the Intel VMD architecture, visit:

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Intel VMD / VROC architecture: Software view of the technology

From the software perspective, in addition to the CPU-integrated hardware functions, software drivers and utilities are implemented to support the Intel VMD and VROC functions.

Firstly, there is VROC firmware driver within the UEFI to handle the pre-boot RAID management functions as needed. This driver allows the OS to be installed on and boot from an Intel VROC RAID storage devices.

Secondly, there are several OS kernel space drivers needed to be implemented to support the media access, as well as RAID functions for the VROC. These include NVMe driver, VMD (Volume Management Device) driver and MD (multiple devices) driver:

- NVMe driver essentially implements the media administration and I/O access to the NVMe SSD storage devices.
- VMD driver works with the VMD hardware controller inside the CPU PCIe root complex, which aggregates the attached NVMe SSD storage devices and maps the entire PCIe sub trees (child devices) into its own address space and enables more or less virtual HBAs for the connected NVMe SSDs.
- MD driver is the driver to perform the VROC RAID function.

Thirdly, there are OS user space drivers and utilities needed to be implemented to support the RAID administration and management functions. These include mdadm and ledmon utilities:

- mdadm is the primary software to configure and manage Intel VROC RAID, which is a native Linux tool upgraded to support Intel VROC RAID functions.

- ledmon is a daemon process used to monitor the state of MD RAID devices or the state of block devices.

The relationship among the hardware and all the above-mentioned drivers and utilities are summarized in Table 1.

Table 1 Hardware and software layers

Layer	Driver
OS User Space	mdadm / ledmon (VROC)
OS Kernel Space	md (VROC) vmd nvme
UEFI	VROC FW driver
Hardware	VMD in the CPU NVMe SSD

Intel VROC NVMe RAID support on ThinkSystem servers

Intel VROC NVMe RAID is supported on

- ThinkSystem V2 servers with Gen 3 Intel Xeon Scalable processors (Ice Lake or Cedar Island).
- Selected ThinkSystem V1 servers with Gen 1 or Gen 2 Xeon Scalable processors, including SR650 and SR630

By default, VROC NVMe RAID support is limited to use with only Intel-branded NVMe drives. If you wish to enable RAID support for non-Intel NVMe SSDs, select the VROC Premium license using the ordering information in the following table. VROC Premium is fulfilled as a Feature on Demand (FoD) license and is activated via the XCC management processor user interface.

Table 2 Intel VROC Premium license

Part number	Feature code	Description
4L47A39164	B96G	Intel VROC (VMD NVMe RAID) Premium

VROC Premium is only needed for non-Intel NVMe drives in a RAID configuration. You do not need the VROC Premium license upgrade under any of the following conditions:

- If you have Intel NVMe drives connected to the onboard NVMe ports, you do not need VROC Premium
- If you have non-Intel NVMe drives connected to the onboard NVMe ports, but you don’t require RAID support, you do not need VROC Premium

VROC Premium license activation key is needed for non-Intel NVMe drives in a RAID configuration. You can include the feature code listed in Table 2 in your server order if you are using the configure-to-order process to purchase a server. The server will then arrive with the VROC Premium license already installed and activated.

Alternatively, purchase the part number listed in Table 2 and You will receive an email with instructions on how to obtain the activation key:

1. Go to the Lenovo Key Management System web site, https://fod.lenovo.com
2. Click **Help** and click the link for **instructions on how to redeem your authorization codes**.
3. Review the document and follow the section "How to Redeem Authorization Code" to get the VROC Premium activation key from Lenovo.
4. Once the activation key file is received from Lenovo, open the web interface to the XClarity Controller of the server with the non-Intel NVMe drives.
5. Log in using your XCC credentials and click **BMC Configuration** → **Licence** and click the Upgrade Licence button as shown in Figure 2.

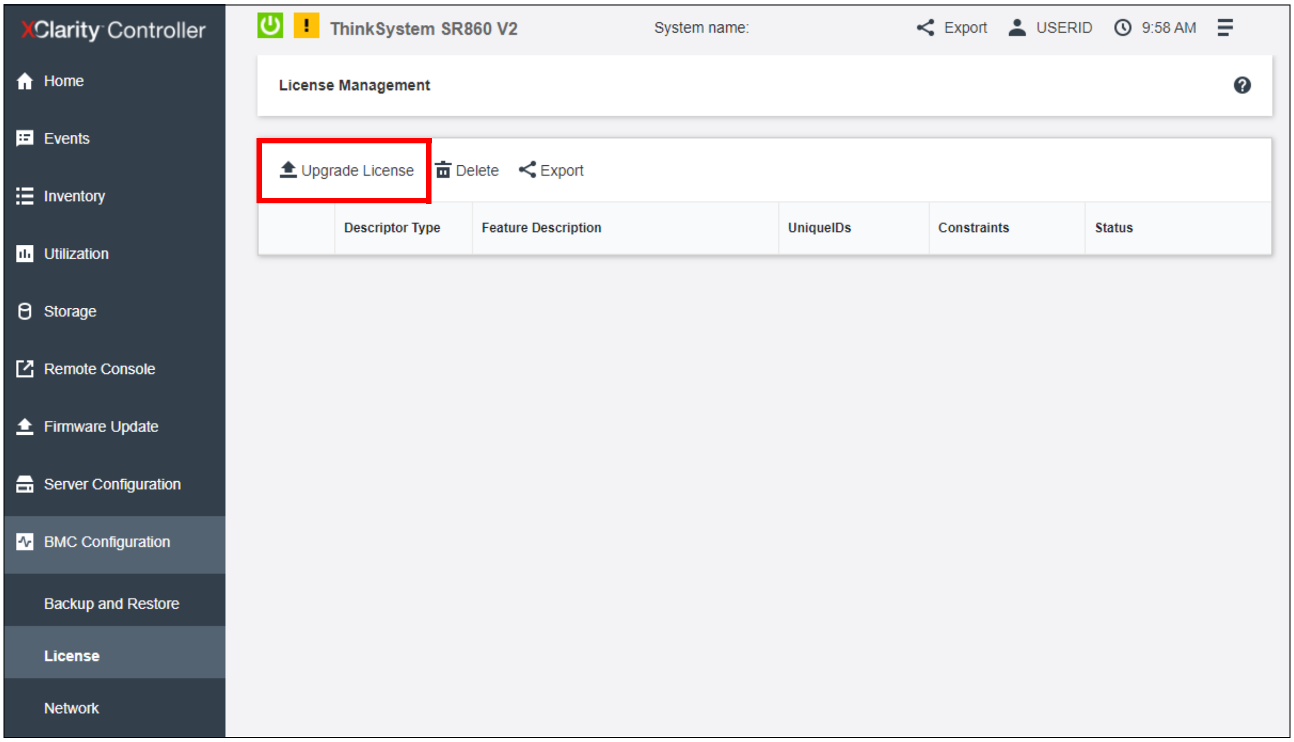


Figure 2 Adding the VROC Premium license to XCC

6. Click **Browse** to locate the activation key you downloaded from the LKMS web site then click **Import** as shown in Figure 3.

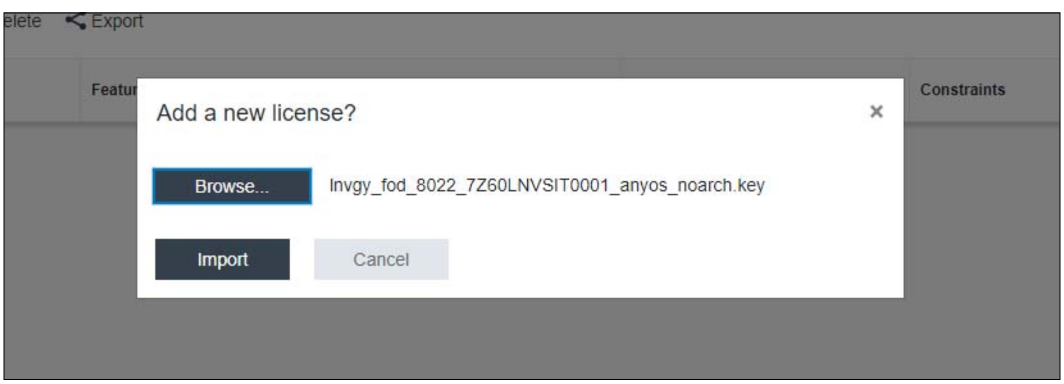
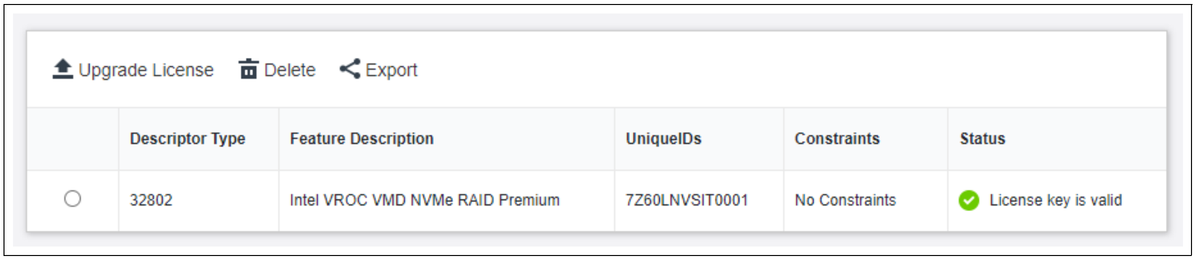


Figure 3 Add a new license

The License Management screen now shows the license as shown in Figure 4. You can now configure VROC RAID using non-Intel NVMe drives.



Upgrade License Delete Export

	Descriptor Type	Feature Description	UniqueIDs	Constraints	Status
○	32802	Intel VROC VMD NVMe RAID Premium	7Z60LNVSIT0001	No Constraints	License key is valid

Figure 4 VROC Premium license applied

Implementing Intel VMD and Intel VROC NVMe

In this section, we will demonstrate how the Intel VMD and VROC can be used with Lenovo ThinkSystem SR860 V2 server which is based on the 3rd Gen Intel Xeon Scalable processors.

Figure 5 ThinkSystem SR860 V2

The following operating systems are the minimum OS versions supported on the SR860 V2 that also support Intel VMD technology:

- Windows Server 2016 or later
- Red Hat Enterprise Linux 7.9 or later
- SUSE Linux Enterprise Server 12 SP5 or later
- VMware ESXi 7.0 U1 or later

With the Windows and Linux OSes, VMD + VROC support the following RAID levels (Non-Intel-branded NVMe SSDs will require the Intel VROC Premium license applied to the server)

- RAID 0
- RAID 1
- RAID 5
- RAID 10

With VMware ESXi, RAID support is limited to RAID 1 only.

The ThinkSystem SR860 V2 server has 48x 2.5-inch hot-swap drive bays that are accessible from the front of the server. Depending on server configuration and installed backplanes, these bays are connected either to the integrated 6 Gbps SATA controller, one or more SAS/SATA RAID controllers or directly to PCIe lanes for NVMe drives.

In our test environment, we installed four Intel P4610 PCIe NVMe 3.2 TB SSDs directly to onboard PCIe connectors. Since these are Intel NVMe drives, the Intel VROC Premium license key was not needed.

Configuring UEFI to enable VMD

The steps to configure the server UEFI to enable the VMD feature are as follows.

1. Press F1 when prompted during server boot to enter System Setup
2. Select **System Settings** → **Devices and I/O Ports** → **Intel VMD Technology**. Figure 6 appears.

Figure 6 Intel VMD Technology screen in UEFI

3. Press Enter to Enable VMD if it isn't already.
4. Save the settings and reboot the system
5. When prompted, press F1 to reenter System Setup.
6. Select System Settings > Storage. Here you should see the installed NVMe SSDs as well as an entry for Intel VROC, Figure 7.

Figure 7 Storage screen in UEFI

This confirms that Intel VROC is enabled and ready to configure a RAID volume.

Creating a RAID volume in UEFI

The steps to create a RAID volume in UEFI using Intel VROC are as follows.

1. In UEFI, select **System Settings** → **Storage** → **Intel Virtual RAID on CPU**. Figure 8 appears.

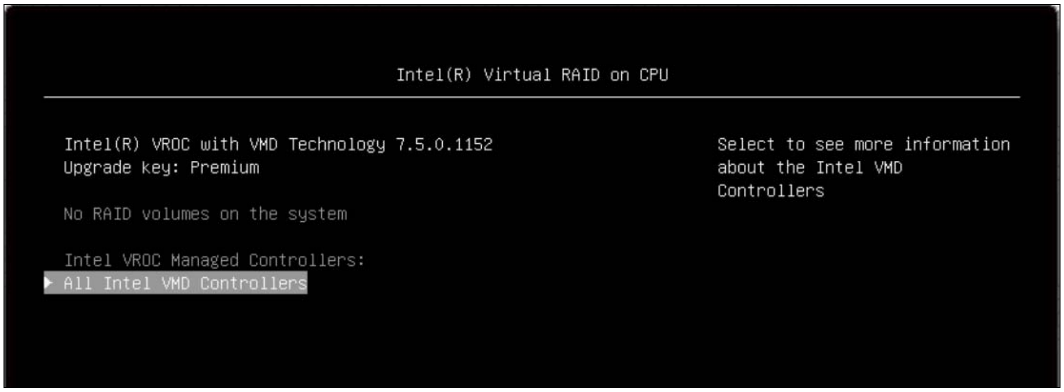


Figure 8   Intel VROC page in System Setup

2. Select **All Intel VMD Controller**. Figure 9 appears.

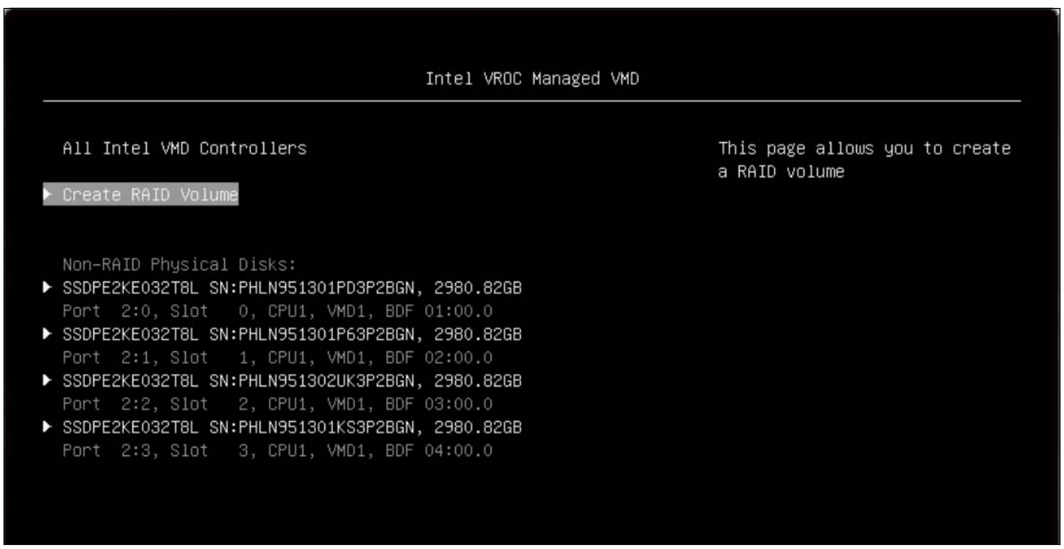


Figure 9   Intel VROC Managed VMD page

3. Select **Create RAID Volume**.
4. For RAID Level, select **RAID5 (Parity)** as shown in Figure 10.

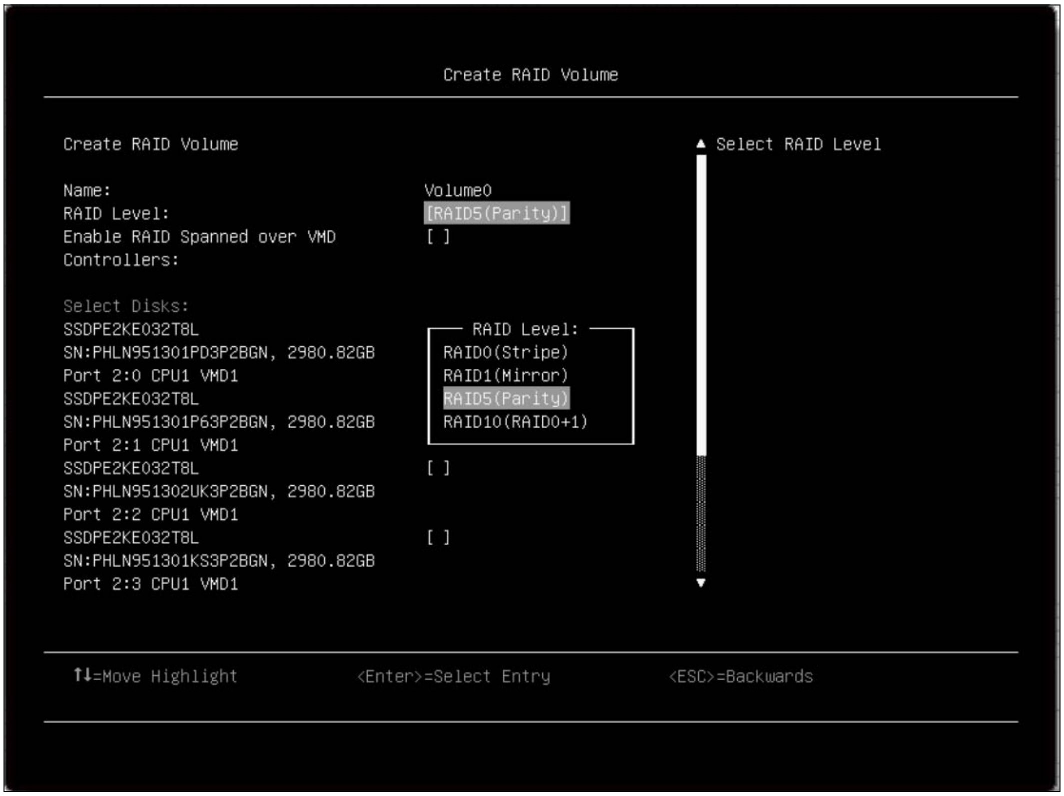


Figure 10 Create RAID volume

5. Leave **Enable RADI Spanned over VVMD** unselected. We are installing a bootable OS on this RAID volume and this option is for Data RAID only. OS boot from such RAID is not supported.
6. Select all disks.

7. Scroll down and select **Create Volume** as shown in Figure 11 to make the configuration take effect.

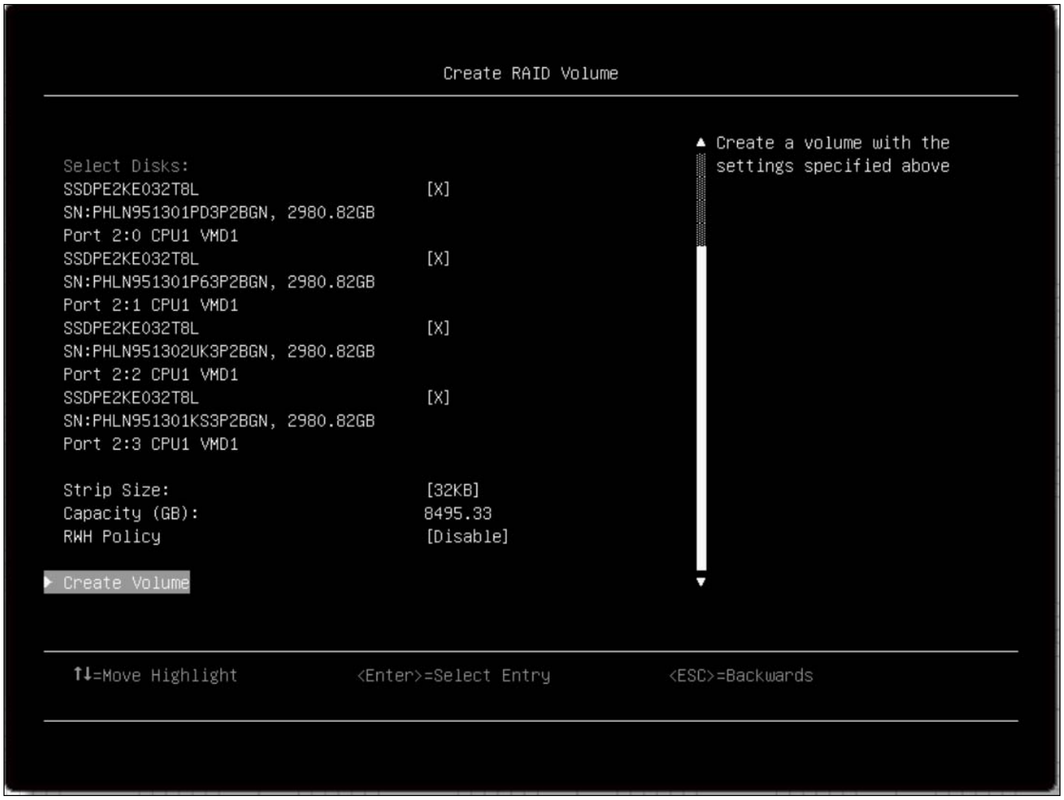


Figure 11 Create Volume

8. The RAID Volume has now been created as Volume0 as shown in Figure 12.

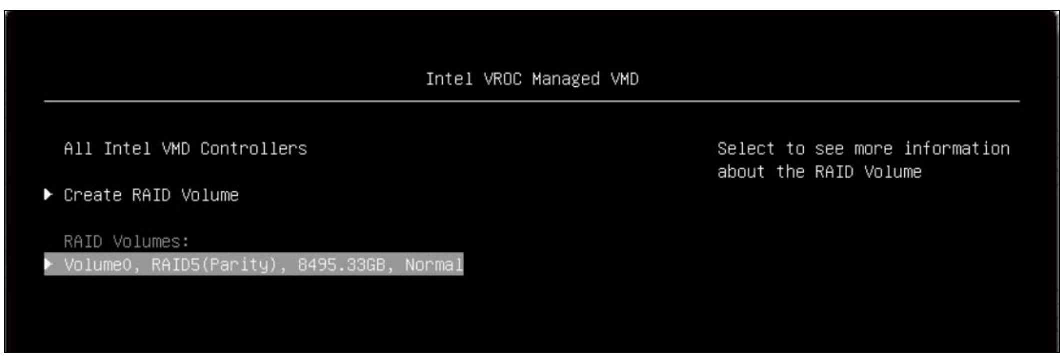


Figure 12 Volume created

9. You can verify this by checking its details as shown in Figure 13.

Figure 13 Volume details

10. Save the settings and reboot the system.

At this point, we can start to install RHEL 8.2 on the newly-created RAID-5 volume.

After starting the OS install, in the Installation Destination screen, select the **BIOS RAID set (raid5)** as the target as shown in Figure 14, and follow the installation instructions to install the operating system.

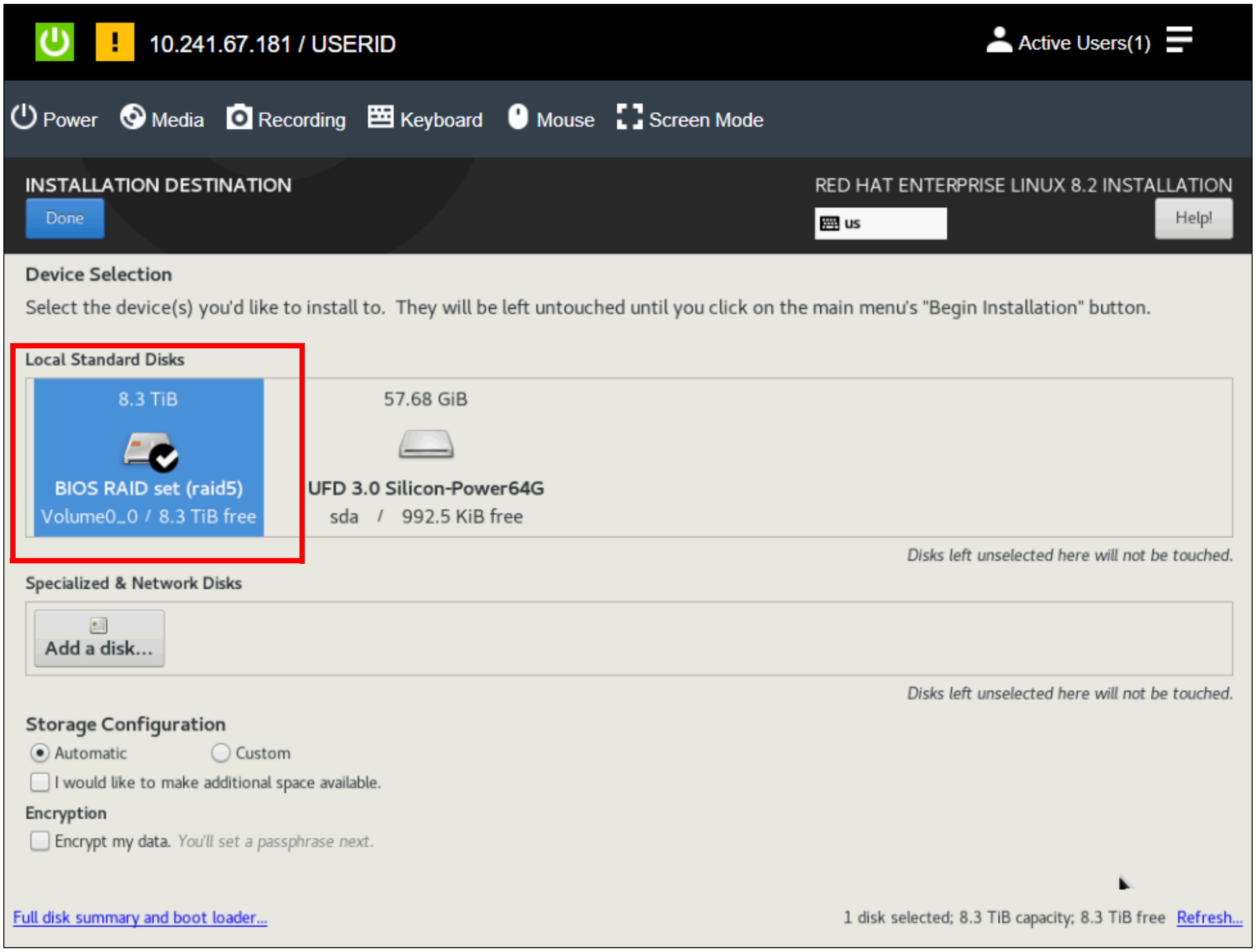


Figure 14 Selecting the NVMe RAID volume as the installation drive

Verify the VMD and VROC functions in the OS

After the RHEL 8.2 OS installation, boot the OS, and do the following:

1. Check the NVMe disks existence with following commands:

   ```
   fdisk -l  | grep nvme
   ls -l /dev/nvme*
   lsblk | grep nvme
   ```

   Output is shown in Example 1.

   Example 1 Commands to verify existences of NVMe drives

   ```
   [root@localhost ~]# fdisk -l  | grep nvme
   Disk /dev/nvme0n1: 2.9 TiB, 3200631791616 bytes, 6251233968 sectors
   Disk /dev/nvme3n1: 2.9 TiB, 3200631791616 bytes, 6251233968 sectors
   Disk /dev/nvme2n1: 2.9 TiB, 3200631791616 bytes, 6251233968 sectors
   Disk /dev/nvme1n1: 2.9 TiB, 3200631791616 bytes, 6251233968 sectors
   [root@localhost ~]# ls -l /dev/nvme*
   ```

```
crw-------. 1 root root 242, 0 Oct 29 05:25 /dev/nvme0
brw-rw----. 1 root disk 259, 0 Oct 29 05:25 /dev/nvme0n1
crw-------. 1 root root 242, 1 Oct 29 05:25 /dev/nvme1
brw-rw----. 1 root disk 259, 3 Oct 29 05:25 /dev/nvme1n1
crw-------. 1 root root 242, 2 Oct 29 05:25 /dev/nvme2
brw-rw----. 1 root disk 259, 2 Oct 29 05:25 /dev/nvme2n1
crw-------. 1 root root 242, 3 Oct 29 05:25 /dev/nvme3
brw-rw----. 1 root disk 259, 1 Oct 29 05:25 /dev/nvme3n1
[root@localhost ~]# lspci | grep -i non-vol
10000:01:00.0 Non-Volatile memory controller: Intel Corporation NVMe Datacenter
SSD [3DNAND, Beta Rock Controller]
10000:02:00.0 Non-Volatile memory controller: Intel Corporation NVMe Datacenter
SSD [3DNAND, Beta Rock Controller]
10000:03:00.0 Non-Volatile memory controller: Intel Corporation NVMe Datacenter
SSD [3DNAND, Beta Rock Controller]
10000:04:00.0 Non-Volatile memory controller: Intel Corporation NVMe Datacenter
SSD [3DNAND, Beta Rock Controller]
[root@localhost ~]# lsblk | grep nvme
nvme0n1         259:0    0   2.9T  0 disk
nvme3n1         259:1    0   2.9T  0 disk
nvme2n1         259:2    0   2.9T  0 disk
nvme1n1         259:3    0   2.9T  0 disk
```

2. Check the VMD enablement with the following commands:

   lspci -tv
   ls -al /sys/block
   dmesg | grep -i vmd

   Output is shown in Example 2. Note the following:

   – The **lspci** command shows that Domain 10000:00 is added by VMD in PCI buses information.

   – The **ls /sys/block** command output shows that 0000:39:05.5 is VMD device and pci10000:00 is PCIe domain added by VMD)

   – The dmesg command allows you to check VMD device messages in the OS log

Example 2 Commands to check that VMD has been enabled

```
[root@localhost block]# lspci -tv
+-[10000:00]-+-00.0-[01]----00.0  Intel Corporation NVMe Datacenter SSD
[3DNAND, Beta Rock Controller]
 |           +-01.0-[02]----00.0  Intel Corporation NVMe Datacenter SSD
[3DNAND, Beta Rock Controller]
 |           +-02.0-[03]----00.0  Intel Corporation NVMe Datacenter SSD
[3DNAND, Beta Rock Controller]
 |           \-03.0-[04]----00.0  Intel Corporation NVMe Datacenter SSD
[3DNAND, Beta Rock Controller]
... (skipped)

[root@localhost block]# ls -al /sys/block
total 0
drwxr-xr-x.  2 root root 0 Oct 29 05:25 .
dr-xr-xr-x. 13 root root 0 Oct 29 05:25 ..
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 dm-0 -> ../devices/virtual/block/dm-0
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 dm-1 -> ../devices/virtual/block/dm-1
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 dm-2 -> ../devices/virtual/block/dm-2
```

```
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 md126 -> ../devices/virtual/block/md126
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 md127 -> ../devices/virtual/block/md127
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 nvme0n1 ->
../devices/pci0000:39/0000:39:05.5/pci10000:00/10000:00:00.0/10000:01:00.0/nvme
/nvme0/nvme0n1
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 nvme1n1 ->
../devices/pci0000:39/0000:39:05.5/pci10000:00/10000:00:01.0/10000:02:00.0/nvme
/nvme1/nvme1n1
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 nvme2n1 ->
../devices/pci0000:39/0000:39:05.5/pci10000:00/10000:00:02.0/10000:03:00.0/nvme
/nvme2/nvme2n1
lrwxrwxrwx.  1 root root 0 Oct 29 05:25 nvme3n1 ->
../devices/pci0000:39/0000:39:05.5/pci10000:00/10000:00:03.0/10000:04:00.0/nvme
/nvme3/nvme3n1

[root@localhost block]# dmesg | grep -i vmd
[    2.289217] vmd 0000:39:05.5: PCI host bridge to bus 10000:00
[    2.291104] vmd 0000:39:05.5: Bound to PCI domain 10000
[    2.294872] vmd 0000:84:05.5: PCI host bridge to bus 10001:00
[    2.297417] vmd 0000:84:05.5: Bound to PCI domain 10001
... (skipped)
```

Additional RAID function

The above instructions have demonstrated that the OS can be installed on the RAID volume that is previously created within the UEFI after enabling the VMD feature.

For other functions such as detailed RAID information, the creation of a new RAID within the OS, or the RAID volumes management etc, use the mdadm command line utility, which supports the intel Matrix Storage Manager (IMSM) metadata format. See Figure 15.

```
[root@localhost ~]# mdadm --detail-platform
mdadm: imsm capabilities not found for controller: /sys/devices/pci0000:00/0000:00:17.0 (type SATA)
mdadm: imsm capabilities not found for controller: /sys/devices/pci0000:00/0000:00:11.5 (type SATA)
       Platform : Intel(R) Virtual RAID on CPU
        Version : 7.5.0.1152
    RAID Levels : raid0 raid1 raid10 raid5
    Chunk Sizes : 4k 8k 16k 32k 64k 128k
    2TB volumes : supported
      2TB disks : supported
      Max Disks : 48
    Max Volumes : 2 per array, 24 per controller
 3rd party NVMe : supported
 I/O Controller : /sys/devices/pci0000:39/0000:39:05.5 (VMD)
 NVMe under VMD : /dev/nvme0n1 (PHLN951301PD3P2BGN)
 NVMe under VMD : /dev/nvme3n1 (PHLN951301KS3P2BGN)
 NVMe under VMD : /dev/nvme2n1 (PHLN951302UK3P2BGN)
 NVMe under VMD : /dev/nvme1n1 (PHLN951301P63P2BGN)
 I/O Controller : /sys/devices/pci0000:84/0000:84:05.5 (VMD)

[root@localhost ~]#
```

Figure 15   mdadm command

For more details about the usage of mdadm command line utility, please refer to the intel document which can be found on the web:

- Intel VROC Linux Software User Guide

  https://www.intel.com/content/dam/support/us/en/documents/memory-and-storage/ssd-software/Linux_VROC_6-0_User_Guide.pdf

Change history

October 25, 2021

- ThinkSystem SR650 and SR630 now support Intel VROC NVMe RAID - “Intel VROC NVMe RAID support on ThinkSystem servers” on page 5
- Clarified the RAID levels supported - “Implementing Intel VMD and Intel VROC NVMe” on page 7

First published: May 17, 2021

Author

Kelvin Shieh is the OS Development Technical Lead for the Lenovo Infrastructure Solutions Group, based in Taipei, Taiwan.

Thanks to the following people for their contributions to this project:

- David Watts, Lenovo Press

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on October 27, 2021.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1466

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available from https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®	Lenovo(logo)®	ThinkSystem™

The following terms are trademarks of other companies:

Intel, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Hyper-V, Windows, Windows Server, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.