

Technical Report

ONTAP Select with NVMe Storage

Architecture and Configuration Guide

Arvind Jagannath, NetApp September 2017 | TR-4625

Abstract

This report provides an architecture overview and serves as a quick start guide to using NVMe storage behind NetApp® ONTAP® Select in a KVM hypervisor environment.



TABLE OF CONTENTS

1	Intr	oduction	3		
2	Arc	Architecture			
	2.1	NVMe Storage for ONTAP Select	3		
	2.2	Storage Configuration Overview	4		
	2.3	Required Packages to Support NVMe and MD-RAID	5		
3	Configuration				
	Step	o 1: Creating a RAID-6 Array Using NVMe Disks	6		
	Step	2: Initializing the New RAID Array and Checking the Status	7		
	Step	3: Making the New RAID Array Persistent Across Host Reboots	8		
	Step	4: Creating Storage Pools for ONTAP Select Consumption	8		
4	Troubleshooting				
	4.1	Physical Volume (NVMe Devices) Disappearing After a Reboot	9		
	4.2	NVMe Disk Failure on Boot	9		
Re	ferer	1Ces	12		
Ve	rsior	n History	12		
LIS	OF F	FIGURES			
Fig	ure 1	ONTAP Select with LVM on a host with a hardware RAID controller.	4		
Fig	ure 2	ONTAP Select with LVM on a host with the MD-RAID software RAID layer	4		
Fig	ure 3) MD-RAID software RAID layer configuration.	5		

1 Introduction

NetApp® ONTAP® Select is a software-defined storage solution providing industry-leading enterprise data management capabilities. ONTAP Select helps accelerate deployment with cloud-like agility and increased flexibility on your choice of commodity servers and hypervisors.

Nonvolatile Memory Express (NVMe) enables a scalable, high-performance architecture designed for enterprise server systems utilizing solid-state drive (SSD) storage technology with Peripheral Component Interconnect Express (PCIe) Gen 3 interconnect. The technology was developed to reduce latency and provide faster CPU-to-data-storage-device performance for advanced computing. NVMe-based storage enables applications such as real-time analytics, transactional databases, and financial services.

NVMe-based storage improves the capabilities of ONTAP Select by providing a new data storage option. The storage administrator must set up NVMe-based storage manually before deploying ONTAP Select. NVMe support requires a software RAID controller on Linux and is thus currently supported only for ONTAP Select on the Kernel-Based Virtual Machine (KVM) hypervisor. This document describes the architecture and serves as a guide for quickly configuring, installing, and deploying NVMe as the backend storage for ONTAP Select on KVM.

2 Architecture

This section describes the basic architecture for using NVMe devices with ONTAP Select, and illustrates the layering changes required to support this new storage option.

2.1 NVMe Storage for ONTAP Select

Traditionally, NetApp ONTAP Select has supported SAS, NL-SAS, SATA, or SSD drives connected as either direct-attached storage (DAS) or external storage arrays. For direct-attached storage, a hardware RAID controller installed on the host hardware protects the underlying storage against individual drive failures and provides RAID groups. The host and hypervisor then help abstract these RAID groups into a consistent and unified namespace such as a LUN or a Virtual Machine Disk (VMDK).

The newer NVMe storage introduces a new interface that doesn't support such abstractions, because the PCIe interfaces that NVMe drives use do not currently support hardware RAID controllers. Instead of hardware RAID, a software RAID mechanism provided by Linux is used.

2.2 Storage Configuration Overview

Figure 1) ONTAP Select with LVM on a host with a hardware RAID controller.

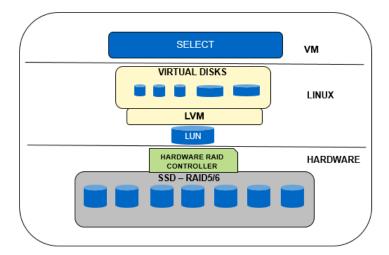
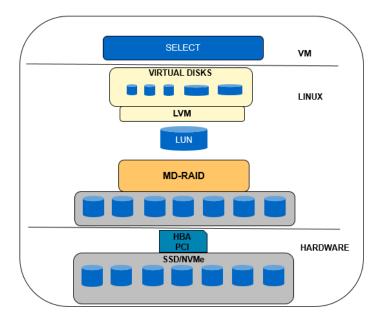


Figure 2) ONTAP Select with LVM on a host with the MD-RAID software RAID layer.



In the absence of the hardware RAID controller, a software RAID layer (MD-RAID) is required to perform the required RAID functionality. The MD-RAID layer (along with the mdadm utility) provides the software RAID functionality. Software RAID configurations can include anything presented to the Linux kernel as a block device. This includes whole hard drives (for example, /dev/sda) or their partitions.

As shown in Figure 2, the NVMe devices seen as individual block devices in Linux can be combined to form RAID groups using the MD-RAID layer and exposed as a single LUN to the logical volume manager (LVM) layer. LVM could then be used to carve up multiple logical volumes to be exposed to the virtual machine (VM). This final step would be similar to a traditional configuration with a hardware RAID controller.

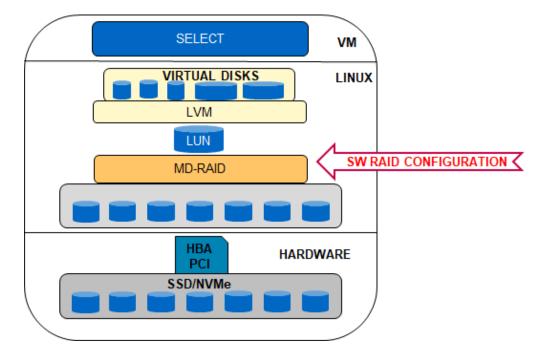
2.3 Required Packages to Support NVMe and MD-RAID

Linux OS/Package and Version	Description
nvme-cli-1.1-1.fc25.x86_64	NVMe user space tooling for Linux. This powerful command- line utility is used to report statistics and the characteristics of the NVMe device. It can also run diagnostic commands.
mdadm-3.4-14.e17.x86_64	The mdadm program is used to create, manage, and monitor Linux MD (software RAID) devices. It provides similar functionality to the raidtools package. However, mdadm is a single program, and it can perform almost all functions without a configuration file, though a configuration file is used to help with some common tasks.

3 Configuration

This section describes how to set up a new software RAID group using MD-RAID as shown in Figure 3. This section shows how the storage administrator can create a RAID array of NVMe devices using the mdadm utility, check the status of the initialization, make the RAID array persistent, and finally use the array to create storage pools for NetApp ONTAP Select.

Figure 3) MD-RAID software RAID layer configuration.



Step 1: Creating a RAID-6 Array Using NVMe Disks

In this example, there are 24 NVMe devices on this system: /dev/nvme0n1 through /dev/nvme23n1. Each NVMe drive is a PCIe device directly attached to the host PCIe bus.

List all the available NVMe devices in the system.

This listing helps you decide which NVMe devices can belong to the software RAID group.

```
[root@rhel7host ~]# nvme list
Node SN Model Namespace Usage Format FW Rev
._____
/dev/nvme0n1 S23YNAAH200466 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
/dev/nvme10n1 S23YNAAH200400 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme11n1 S23YNAAH200471 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nyme12n1 S23YNAAH200507 SAMSUNG MZOLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
BXV54010
BXV54010
/dev/nvme15n1 S23YNAAH200401 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
BXV54010
/dev/nvme17n1 S23YNAAH200402 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme18n1 S23YNAAH200488 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nyme19n1 $23YNAAH200491 $AM$UNG MZOLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme1n1 S23YNAAH200468 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme20n1 S23YNAAH200406 SAMSUNG MZOLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
BXV54010
BXV54010
/dev/nvme23n1 S23YNAAH200487 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme2n1 S23YNAAH200464 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme3n1 S23YNAAH200465 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
/dev/nvme4n1 S23YNAAH200213 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
BXV54010
/dev/nvme6n1 S23YNAAH200575 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme7n1 S23YNAAH200572 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme8n1 S23YNAAH200565 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV54010
/dev/nvme9n1 S23YNAAH200508 SAMSUNG MZQLV480HCGR-00003 1 480.10 GB / 480.10 GB 512 B + 0 B
BXV5401Q
```

Create a software RAID group using NVMe devices.

The mdadm command is used to combine NVMe drives into a single LUN. MD-RAID lets you designate a few drives as spare drives to use as replacements when a drive completely fails. Note that mdadm automatically uses the spares, and no manual intervention is necessary. The following command creates a RAID-6 array called /dev/md0 using 22 disks, with the remaining 2 marked as hot spares.

```
[root@rhel7host ~]# mdadm --create --verbose /dev/md0 --level=6 --raid-devices=22 /dev/nvme0n1 /dev/nvme1n1 /dev/nvme2n1 /dev/nvme3n1 /dev/nvme4n1 /dev/nvme5n1 /dev/nvme6n1 /dev/nvme7n1 /dev/nvme8n1 /dev/nvme9n1 /dev/nvme10n1 /dev/nvme1ln1 /dev/nvme12n1 /dev/nvme13n1 /dev/nvme14n1 /dev/nvme15n1 /dev/nvme16n1 /dev/nvme17n1 /dev/nvme18n1 /dev/nvme19n1 /dev/nvme20n1 /dev/nvme21n1 --spare-devices=2 /dev/nvme22n1 /dev/nvme23n1 mdadm: layout defaults to left-symmetric mdadm: layout defaults to left-symmetric mdadm: chunk size defaults to 512K mdadm: size set to 468720128K mdadm: size set to 468720128K mdadm: automatically enabling write-intent bitmap on large array mdadm: Defaulting to version 1.2 metadata mdadm: array /dev/md0 started.
```

Step 2: Initializing the New RAID Array and Checking the Status

View details of the new RAID group.

You can view details of the new array by using the --detail argument:

```
[root@rhel7host ~]# mdadm --detail /dev/md0
/dev/md0:
Version: 1.2
Creation Time : Tue Aug 22 04:54:50 2017
Raid Level : raid6
Array Size: 9374402560 (8940.13 GiB 9599.39 GB)
Used Dev Size: 468720128 (447.01 GiB 479.97 GB)
Raid Devices: 22
Total Devices: 24
Persistence : Superblock is persistent
Intent Bitmap : Internal
Update Time: Tue Aug 22 04:54:56 2017
State : clean, resyncing
Active Devices: 22
Working Devices: 24
Failed Devices : 0
Spare Devices : 2
Layout : left-symmetric
Chunk Size : 512K
Resync Status: 0% complete
Name : rhel7host:0 (local to host rhel7host)
UUID: 77922928:d1d4d0c5:fdccf295:f3376948
Events : 1
Number Major Minor RaidDevice State
0 259 0 0 active sync /dev/nvme0n1
1 259 2 1 active sync /dev/nvme1n1
2 259 1 2 active sync /dev/nvme2n1
3 259 3 3 active sync /dev/nvme3n1
4 259 4 4 active sync /dev/nvme4n1
5 259 6 5 active sync /\text{dev/nvme}5\text{n}1
 6 259 5 6 active sync /dev/nvme6n1
 7 259 7 7 active sync /dev/nvme7n1
8 259 9 8 active sync /dev/nvme8n1
9 259 8 9 active sync /dev/nvme9n1
10 259 10 10 active sync /dev/nvme10n1
11 259 11 11 active sync /dev/nvme11n1
12 259 12 12 active sync /dev/nvme12n1
13 259 13 13 active sync /dev/nvme13n1
14 259 14 14 active sync /dev/nvme14n1
15 259 15 15 active sync /dev/nvme15n1
16 259 16 16 active sync /dev/nvme16n1
17 259 18 17 active sync /dev/nvme17n1
18 259 17 18 active sync /dev/nvme18n1
19 259 19 19 active sync /dev/nvme19n1
```

```
20 259 20 20 active sync /dev/nvme20n1
21 259 21 21 active sync /dev/nvme21n1
22 259 22 - spare /dev/nvme22n1
23 259 23 - spare /dev/nvme23n1
```

View status of the new RAID group to determine the progress of the sync.

You can view the status of the new array (the initial sync) by looking at /proc/mdstat.

For an array of this size, the sync takes about 90 minutes to complete. NetApp strongly recommends that you complete this step before making use of the new device /dev/md0.

Verify that the status of the sync is complete.

The status is shown when the sync is complete.

```
[root@rhel7host ~]# cat /proc/mdstat
Personalities : [raid6] [raid5] [raid4]
md0 : active raid6 nvme23n1[23](S) nvme22n1[22](S) nvme21n1[21] nvme20n1[20] nvme19n1[19]
nvme18n1[18] nvme17n1[17] nvme16n1[16] nvme15n1[15] nvme14n1[14] nvme13n1[13] nvme12n1[12]
nvme11n1[11] nvme10n1[10] nvme9n1[9] nvme8n1[8] nvme7n1[7] nvme6n1[6] nvme5n1[5] nvme4n1[4]
nvme3n1[3] nvme2n1[2] nvme1n1[1] nvme0n1[0]
9374402560 blocks super 1.2 level 6, 512k chunk, algorithm 2 [22/22] [UUUUUUUUUUUUUUUUUUUUUU
bitmap: 0/4 pages [0KB], 65536KB chunk
unused devices: <none>
```

Verify that a daemon corresponding to the new RAID array is running.

A daemon dedicated to this new RAID array (/dev/md0) is also automatically created.

```
[root@rhel7host ~] # ps auxwww | grep md0
root 5301 25.8 0.0 0 0 ? S 05:24 164:08 [md0_raid6]
```

Step 3: Making the New RAID Array Persistent Across Host Reboots

Save the RAID configuration to the configuration file.

The array configuration for the new RAID array (device /dev/md0) should be written to a configuration file so that it can be reassembled on a server restart.

On a RHEL7 host, append the details for the scan into the mdadm.conf file.

```
[root@rhel7host ~]# mdadm --detail --scan >> /etc/mdadm.conf
[root@rhel7host ~]# cat /etc/mdadm.conf
ARRAY /dev/md0 metadata=1.2 spares=2 name=rhel7host:0 UUID=77922928:d1d4d0c5:fdccf295:f3376948
```

Warning: Failure to complete this step will cause your NVMe RAID array to be unavailable the next time the host is rebooted. The automatic restarting of the ONTAP Select VM depends on storage pools, which in turn depend on the RAID array being present after the reboot.

Step 4: Creating Storage Pools for ONTAP Select Consumption

Create and start storage pools using the newly created RAID groups to use with ONTAP Select.

Storage pools can then be created as before (also described in <u>TR-4613</u>) and can be consumed by ONTAP Select using the new /dev/md0 device. This is the similar to the way storage pools can be created and consumed using LUNs provided by hardware RAID.

```
[root@rhel7host ~]# virsh pool-define-as swraid logical --source-dev /dev/md0 -target=/dev/swraid
Pool swraid defined
[root@rhel7host ~]# virsh pool-build swraid
Pool swraid built
[root@rhel7host ~]# virsh pool-start swraid
Pool swraid started
[root@rhel7host ~]# virsh pool-autostart swraid
Pool swraid marked as autostarted
```

4 Troubleshooting

4.1 Physical Volume (NVMe Devices) Disappearing After a Reboot

When a host is rebooted, the physical volume (PV) can disappear. This causes the storage pool or volume group to be unavailable after the reboot.

There is a recent Red Hat bug related to this issue:

Bug 1445924 - Volume groups not activating due to mdadm devices reporting degraded

Workaround

Disable lymetad, so that it falls back to legacy LVM activation. In the LVM configuration file, /etc/lym.conf, set use lymetad to 0.

4.2 NVMe Disk Failure on Boot

In simulating disk failures with a physical pull of an NVMe device in the RAID group, tests showed that the volume group might not become available in time for the storage pool to restart automatically. This problem might prevent the VMs from automatically restarting.

For this test, the drive was pulled out after the host was powered down, and then the host was powered back up without the device being plugged back in.

In this situation, when the server boots back up, the /var/log/messages indicate that the volume group (VG) was not available when the storage pool tried to start and the VMs did not autostart, as shown here:

```
Aug 21 13:15:48 rhel7host journal: unsupported configuration: cannot find logical volume group name 'swraid'
Aug 21 13:15:48 rhel7host journal: internal error: Failed to autostart storage pool 'swraid': unsupported configuration: cannot find logical volume group name 'swraid'
Aug 21 13:15:49 rhel7host journal: Cannot access storage file '/dev/swraid/nvme-2nc-01_sdotconfig.iso' (as uid:0, gid:0): No such file or directory
Aug 21 13:15:49 rhel7host journal: internal error: Failed to autostart VM 'nvme-2nc-01': Cannot access storage file '/dev/swraid/nvme-2nc-01_sdotconfig.iso' (as uid:0, gid:0): No such file or directory
```

Workaround

Start the storage pool swraid, and then start the ONTAP VM nvme-2nc-01 manually, as shown here:

```
[root@rhel7host ~]# virsh pool-start swraid
Pool swraid started
[root@rhel7host ~]# virsh pool-list --all
Name State Autostart
```

Once again, you can see the status of the RAID group that is being rebuilt by using the madadm command.

Note that the disk with Number 8 in the output is missing, and is replaced with the spare device /dev/nvme22n1.

```
[root@rhel7host ~] # mdadm --detail /dev/md0
/dev/md0:
Version: 1.2
Creation Time : Tue Aug 22 04:54:50 2017
Raid Level : raid6
Array Size: 9374402560 (8940.13 GiB 9599.39 GB)
Used Dev Size: 468720128 (447.01 GiB 479.97 GB)
Raid Devices : 22
Total Devices: 23
Persistence: Superblock is persistent
Intent Bitmap : Internal
Update Time : Wed Aug 23 05:34:02 2017
State : clean, degraded, recovering
Active Devices: 21
Working Devices: 23
Failed Devices : 0
Spare Devices : 2
Layout : left-symmetric
Chunk Size : 512K
Rebuild Status : 0% complete
Name : rhel7host:0 (local to host rhel7host)
UUID: 77922928:d1d4d0c5:fdccf295:f3376948
Events: 950
Number Major Minor RaidDevice State
0 259 0 0 active sync /dev/nvme0n1
1 259 1 1 active sync /dev/nvme1n1
2 259 2 2 active sync /dev/nvme2n1
3 259 3 3 active sync /dev/nvme3n1
4 259 4 4 active sync /dev/nvme4n1
5 259 5 5 active sync /dev/nvme5n1
6 259 6 6 active sync /dev/nvme6n1
7 259 8 7 active sync /dev/nvme7n1
```

```
23 259 21 8 spare rebuilding /dev/nvme22n1
9 259 9 9 active sync /dev/nvme8n1
10 259 10 10 active sync /dev/nvme9n1
11 259 7 11 active sync /dev/nvme10n1
12 259 19 12 active sync /dev/nvme11n1
13 259 16 13 active sync /dev/nvme12n1
14 259 13 14 active sync /dev/nvme13n1
15 259 11 15 active sync /dev/nvme14n1
16 259 20 16 active sync /dev/nvme15n1
17 259 15 17 active sync /dev/nvme16n1
18 259 14 18 active sync /dev/nvme17n1
19 259 18 19 active sync /dev/nvme18n1
20 259 17 20 active sync /dev/nvme19n1
21 259 12 21 active sync /dev/nvme20n1
22 259 22 - spare /dev/nvme21n1
```

References

- MD-RAID https://en.wikipedia.org/wiki/Mdadm
- Linux RAID wiki https://raid.wiki.kernel.org/index.php/Linux_Raid
- Lenovo white paper on NVMe (contains mdraid information as well) https://lenovopress.com/lp0508.pdf
- Archlinux page on NVMe https://wiki.archlinux.org/index.php/Solid_State_Drives/NVMe
- Red Hat Linux, "Logical Volume Manager Administration"
 https://access.redhat.com/documentation/en-US/Red Hat Enterprise Linux/6/html-single/Logical_Volume_Manager_Administration/index.html
- TR-4613, "ONTAP Select on KVM: Product Architecture and Best Practices" https://www.netapp.com/us/media/tr-4613.pdf

Version History

Version	Date	Document Version History
Version 1.0	September 2017	Initial version

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Copyright © 1994–2017 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.277-7103 (October 1988) and FAR 52-227-19 (June 1987).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

