Rapid-Cold-Start on Power 9 for SAP HANA
Overview and configuration example for a Rapid-Cold-Start solution with NVMe PCIe Cards based on Power 9

Marc Stephan Tauchert

IBM Deutschland Research & Development GmbH
Created on November 06, 2019 – Version 0.0
Last modified on November 14, 2019 – Version 1.0
© Copyright IBM Corporation, 2019
Note: Before using this information and the product it supports, be sure to read the general information under “Copyrights and Trademarks” on page 16 as well as “Disclaimer and Special Notices” on page 16.

© Copyright IBM Corporation 2019. All Rights Reserved. US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp. All trademarks or registered trademarks mentioned herein are the property of their respective holders.
IBM Corporation
Systems and Technology Group
Route 100
Somers, NY 10589
Edition Notice: this is version 1.0 of this document.
Focus: SAP HANA Scale-up and Scale-out solutions
Target:
- SAP HANA 2.0
- POWER8 + POWER9

<table>
<thead>
<tr>
<th>Doc Version</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>Initial Version</td>
</tr>
<tr>
<td>1.0</td>
<td>Publication</td>
</tr>
</tbody>
</table>
Preface

Running SAP HANA on IBM Power Systems offers customers a consistent platform for their HANA-based and traditional applications, best-in-class performance, resilience for critical workloads, and most flexible infrastructure. Existing IT assets - servers, storage, as well as skills and operation procedures - can easily be (re-)used leveraging the SAP HANA Tailored Data Center Integration (TDI) concept, instead of enforcing additional investment into dedicated SAP HANA appliances.

This paper will show how a HANA can be started 7 times faster compared to standard deployments leveraging internal PCIe attached NVMe devices on Power9 Servers.

About This Document

This document is intended for architects and specialists planning an SAP HANA® on POWER® deployment. It describes the design considerations for hardware, networking, and software components of the SAP HANA on POWER solution stack. This guide does not replace existing SAP HANA documentation and sizing guides. It serves as a supplement to the existing SAP HANA documentation and SAP Sizing methods to provide specific guidance on how to meet all SAP requirements when running SAP HANA on IBM Power Systems™, IBM System Storage®, IBM PowerVM®, and Linux Operating System. It describes the requirements for LAN and external SAN topologies. For special topics, own documentation is maintained and referenced at the end.

IBM processes and contacts are introduced helping to obtain a valid configuration based on SAP sizing for SAP HANA.

IBM employees can access the ISICC SAP HANA on Power Systems Community (IBM only) for up-to-date materials complementary to this guide.

The most recent document version can be downloaded from IBM TechDocs:
http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102502

Feel free to provide feedback and change requests for this document via email to isicc@de.ibm.com

About the Author

Marc-Stephan Tauchert is a System Specialist at IBM Germany. In his 25+ years’ experience in designing and implementing SAP Solutions on IBM Power Systems, he has covered various topics and roles. As Technical Sales Consultant and SAP Solution Architect, he supports customer situations regarding the connectivity of SAP Applications and Infrastructure, SAP sizing, and hybrid solutions. He is an expert in SAP Database and Application performance, including SAP HANA.
Table of Contents

Introduction ........................................................................................................................................5
Hardware Planning and Setup ..............................................................................................................5
Implementing Rapid-Cold-Start on Linux on Power ...........................................................................6
Check hardware and software settings ..............................................................................................6
Get multipath devices and setup physical volumes .........................................................................6
Creating volume group and logical volumes ....................................................................................8
Creating software raid on NVMe devices .........................................................................................10
Wedding of NVMe software raid and LVM stripped SVC multipath devices ..................................12
Create filesystem on software raid ..................................................................................................13
Performance comparison between SVC, NVMe only and Rapid-Cold-Start storage environment .....15
Referenced documents .....................................................................................................................15
Copyrights and Trademarks ...............................................................................................................16
Disclaimer and Special Notices .........................................................................................................16

Figures

Figure 1: Rapid-Cold-Start Architecture for Power 9 .......................................................................5
Figure 2: Get the OS release information ..........................................................................................6
Figure 3: Get the linux kernel version ...............................................................................................6
Figure 4: Rescan the san topology to update multipath devices .........................................................7
Figure 5: List multipath information for a specific device ..................................................................8
Figure 6: Create a physical volume ....................................................................................................8
Figure 7: Check physical volume ........................................................................................................8
Figure 9: Creating a volume group .....................................................................................................9
Figure 10: List volume groups ............................................................................................................9
Figure 11: Create a stripped logical volume .......................................................................................9
Figure 12: Check or list logical volumes ............................................................................................9
Figure 13: Detailed view of logical volumes ......................................................................................10
Figure 14: List NVMe PCI adapter .....................................................................................................11
Figure 15: List NVMe PCI adapter details .........................................................................................11
Figure 16: Create Raid 0 software raid via NVMe devices .................................................................11
Figure 17: Check software raid via "mdadm" utility ...........................................................................11
Figure 18: List software raid information by displaying "/proc/mdstat" .............................................12
Figure 19: List NVMe device information's by using "lsblk" .............................................................12
Figure 20: Create a software raid with raid level 1 ..........................................................................13
Figure 21: Displaying technical detailed information on a software raid .........................................13
Figure 22: Create an XFS filesystem on the existing software raid ................................................14
Figure 23: Read Performance comparison .......................................................................................15
Figure 24: Write performance comparison ........................................................................................15
Introduction
This planning guide supports to architecture and implement new technologies on Power for SAP HANA environments. This technical document is based on experimental tests done by the labs of IBM and SAP in Walldorf. This planning guide will focus on the available storage implementation for SAP HANA based on the following products and features:
- Spectrum Virtualize Family
- Internal NVMe PCIe cards

Hardware Planning and Setup
The Figure 1 shows an architecture overview of the Rapid-Cold-Start solution on Power 9 Servers. This solution based on 4 internal, direct attached NVMe PCI adapters and a Spectrum Virtualize Storage System connected via dual VIO’s using nPIV. The SAP HANA Log and SAP HANA Data filesystems are placed as Raid 1 on an intelligent wiring between NVMe PCI adapters and disks from a Spectrum Virtualize. The SAP HANA Log and Data are placed each on a separate Raid 1 software raid. Each Raid 1 array based on 2 disks, where the first disk is a Raid 0 software raid array based on the NVMe PCI devices and the second disk is based on a Spectrum Virtualize solution with LVM stripping. The configuration matches the best practices for SAP HANA xfs filesystems as documented in the SAP Notes.

Figure 1: Rapid-Cold-Start Architecture for Power 9
Implementing Rapid-Cold-Start on Linux on Power

Check hardware and software settings
This chapter will give some hints, how to check and prepare your operating system based on Linux.
Before starting to setup the Rapid-Cold-Start solution, check the OS release (Figure 2) and the version of the installed Linux kernel (Figure 3).

```
linux:~ # cat /etc/os-release
NAME="SLES"
VERSION="12-SP3"
VERSION_ID="12.3"
PRETTY_NAME="SUSE Linux Enterprise Server 12 SP3"
ID="sles"
ANSI_COLOR="0;32"
CPE_NAME="cpe:/o:suse:sles_sap:12:sp3"
```
Figure 2: Get the OS release information

```
linux:~ # uname -a
Linux sleslinux 4.4.176-94.88-default #1 SMP Thu Mar 21 10:52:54 UTC 2019 (dea44ca) ppc64le ppc64le ppc64le
GNU/Linux
```
Figure 3: Get the linux kernel version

Get multipath devices and setup physical volumes
Starting with the configuration of the Spectrum Virtualize devices use the OS utility “rescan-scsi-bus.sh” to make the multipath devices available to the system. The utility will inform about the newly discovered devices (Figure 4). To check and list the multipath devices use the command “multipath”. The option “-ll” will list all multipath devices and their detailed information of used paths. See Figure 5. To view only one device add the device name to the “multipath” command.
Figure 4: Rescan the san topology to update multipath devices

```bash
linux:~ # rescan-scsi-bus.sh -a -m
Scanning SCSI subsystem for new devices
Scanning host 0 for SCSI target IDs 0 1 2 3 4 5 6 7, all LUNs
Scanning for device 1 0 0 0 ...

... OLD: Host: scsi4 Channel: 00 Id: 06 Lun: 34
  Vendor: IBM  Model: 2145  Rev: 0000
  Type: Direct-Access  ANSI SCSI revision: 06
Scanning for device 4 0 6 35 ...
NEW: Host: scsi4 Channel: 00 Id: 06 Lun: 35
  Vendor: IBM  Model: 2145  Rev: 0000
  Type: Direct-Access  ANSI SCSI revision: 06
Attempting to update multipath devices...
Trying to discover new multipath mappings for newly discovered devices...
16 new or changed device(s) found.
  [1:0:0:35]
  [1:0:2:35]
  [1:0:4:35]
  [1:0:6:35]
  [2:0:0:35]
  [2:0:2:35]
  [2:0:4:35]
  [2:0:6:35]
  [3:0:0:35]
  [3:0:2:35]
  [3:0:4:35]
  [3:0:6:35]
  [4:0:0:35]
  [4:0:2:35]
  [4:0:4:35]
  [4:0:6:35]
0 remapped or resized device(s) found.
0 device(s) removed.
```
After discovering all needed multipath devices, each of these devices have to be a physical volume. Run "pvcreate" on the device name to create the physical volume. See Figure 6. Check the created physical volumes with the command “pvs” Figure 7

Creating volume group and logical volumes
The next step is the creation of the volume group and the logical volume. The command “vgcreate” will create the volume group. Only a volume group name and all needed physical volumes are needed. Figure 8 and Figure 9 are an example for creating and displaying a volume group.
On the volume group a stripped logical volume must be created. Therefor the “lvcreate” command is used and the parameter for the amount of used physical volumes, block size for stripping and usage of the volume group is required. The logical volume example in Figure 11 and Figure 11 used a block size of 64 kB and spread over 8 physical volumes. It should use the complete capacity of the volume group. A detailed check of the logical volume and the volume group is possible by using the “lvdisplay” command with the “-m” option and the volume group name (Figure 12).

```
linux:~ # lvcreate -i 8 -I 64 -l 100%VG rcp_hn_dt_vg -n rcp_hn_dt_lv
Logical volume "rcp_hn_dt_lv" created.
```

Figure 10: Create a stripped logical volume

```
linux:~ # lvdisplay
LV     VG     Attr  LSize  Pool Origin Data%  Meta%  Move Log Cpy% Sync Convert
hn_shr_lv  hn_shr_vg -wi-a----- 512.00g
rcp_hn_dt_lvrcp_hn_dt_vg -wi-a----- 6.00t
rcp_hn_lg_lv rcp_hn_lg_vg -wi-a----- 511.97g
root       system -wi-a----- 243.97g
swap       system -wi-a----- 12.00g
```

Figure 11: Check or list logical volumes
Creating the software raid on NVMe devices

The next steps, after creating the logical volumes based on the Spectrum Virtualize infrastructure, are building the software raid based on the attached NVMe devices. First check which NVMe devices are available on the system. To list all NVMe PCI adapters

```bash
linux:~ # lvdisplay rcp_hn_lg_vg -m
--- Logical volume ---
LV Path       /dev/rcp_hn_lg_vg/rcp_hn_lg_lv
LV Name       rcp_hn_lg_lv
VG Name       rcp_hn_lg_vg
LV UUID       iHpow5-zjpP-mUUe-e0rG-eSpK-w5M7-k1TpeW
LV Write Access read/write
LV Creation host, time lsh30041, 2019-10-10 10:12:28 +0200
LV Status     available
# open         0
LV Size        511.97 GiB
Current LE     131064
Segments       1
Allocation     inherit
Read ahead sectors auto
- currently set to 2048
Block device   254:44

--- Segments ---
Logical extents 0 to 131063:
Type           striped
Stripes        8
Stripe size    64.00 KiB
Stripe 0:
  Physical volume /dev/mapper/360050768018087c520000000001757
  Physical extents 0 to 16382
Stripe 1:
  Physical volume /dev/mapper/360050768018087c520000000001758
  Physical extents 0 to 16382
Stripe 2:
  Physical volume /dev/mapper/360050768018087c520000000001759
  Physical extents 0 to 16382
Stripe 3:
  Physical volume /dev/mapper/360050768018087c52000000000175a
  Physical extents 0 to 16382
Stripe 4:
  Physical volume /dev/mapper/360050768018087c52000000000175b
  Physical extents 0 to 16382
Stripe 5:
  Physical volume /dev/mapper/360050768018087c52000000000175c
  Physical extents 0 to 16382
Stripe 6:
  Physical volume /dev/mapper/360050768018087c52000000000175d
  Physical extents 0 to 16382
Stripe 7:
  Physical volume /dev/mapper/360050768018087c52000000000175e
  Physical extents 0 to 16382
```

Figure 12: Detailed view of logical volumes
attached to the system, use the “lsblk” command and grep to the string “^nvme”. This command will list all attached NVMe PCI adapters (Figure 13). To get detailed information’s about a specific NVMe PCI adapter run “lsblk” command followed by the device name. This will show all NVMe partition’s and configured software raids (Figure 14).

```
linux:~ # lsblk | grep ^nvme
nvme0n1 259:0 0 2.9T 0 disk
nvme3n1 259:5 0 2.9T 0 disk
nvme2n1 259:10 0 2.9T 0 disk
nvme1n1 259:15 0 2.9T 0 disk
```

**Figure 13: List NVMe PCI adapter**

```
linux:~ # lsblk /dev/nvme0n1
NAME    MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
nvme0n1  259:0    0   2.9T  0 disk
├─nvme0n1p1 259:20   0 745.2G  0 part
├─nvme0n1p2 259:21   0 745.2G  0 part
└─nvme0n1p3 259:22   0 745.2G  0 part
```

**Figure 14: List NVMe PCI adapter details**

To create a software raid use the “mdadm” utility. Following options are required:
- Number of the software raid
- Raid level
- Number of attached devices
- All device names

For the example solution the software raid for the SAP HANA Log space has 4 NVMe devices and the software raid for the SAP HANA Data space has 12 NVMe devices. See (Figure 15). The NVMe devices are equally distributed over the 4 NVMe PCI adapters to get the best IO performance.

```
linux:~ # mdadm --create 0 --level=0 --raid-devices=4 /dev/nvme0n1p1 /dev/nvme1n1p1 /dev/nvme2n1p1 /dev/nvme3n1p1
mdadm: Defaulting to version 1.2 metadata
mdadm: array /dev/md/0 started.
```

**Figure 15: Create Raid 0 software raid via NVMe devices**

Checking the software raid after creation use the “mdadm” utility or list the content of /proc/mdstat (Figure 17 and Figure 17). The “lsblk” command can be used for getting information about the created software raids (See Figure 18). For detailed information about the array use “mdadm” like the example of Figure 18.

```
linux:~ # mdadm --examine --brief --scan --config=partitions
ARRAY /dev/md/0 metadata=1.2 UUID=90ef4215:39bd631b:4dbcebe7:beb49429 name=lsh3041:0
```

**Figure 16: Check software raid via “mdadm” utility**
Wedding of NVMe software raid and LVM stripped Spectrum Virtualize multipath devices

At this point a software raid 0 and a logical volume stripped over 8 multipath devices exist. To bring both together, a software raid1 will be used. To optimize the read performance (giving preference on read to the software raid of the NVMe devices), set the option “write-mostly” for the logical volume based on the multipath devices from the Spectrum Virtualize. The target size of the new software raid will be determined by the smallest logical storage device. See the example in Figure 19.
Create filesytem on top of the software raid

Last step is to create an XFS filesystem on the software raid1 volumes. Therefore the standard tool “mkfs” is used as demonstrated in Figure 21.
Figure 21: Create an XFS filesystem on the existing software raid

```bash
linux:~ # mkfs -t xfs /dev/md2
meta-data=/dev/md2 isize=256 agcount=16, agsize=8385920 blks
  = sectsz=512 attr=2, projid32bit=1
  = crc=0 finobt=0, sparse=0
data = bsize=4096 blocks=134174720, imaxpct=25
  = sunit=128 swidth=512 blks
naming =version 2 bsize=4096 ascii-ci=0 ftype=1
log =internal log bsize=4096 blocks=65520, version=2
  = sectsz=512 sunit=8 blks, lazy-count=1
realtime =none extsz=4096 blocks=0, rtextents=0
```
Performance comparison between SVC, NVMe only and Rapid-Cold-Start storage environments

The next two Figures below are showing the difference and the impact on the performance between the 3 different storage configurations. The graphs demonstrate the positive effect on the read performance without impacting the write performance.

**Figure 22: Read Performance comparison**

**Figure 23: Write performance comparison**

**Referenced documents**

[SAP HANA on IBM Power Systems and IBM System Storage - Guides](#)
Any references in this information to non-IBM 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 IBM 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 results 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 measurement may have been estimated through extrapolation. Actual results may vary.

Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products.

All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs

ANY INFORMATION HEREIN IS PROVIDED “AS IS” WITHOUT WARRANTY OR INDEMNIFICATION OF ANY KIND BY IBM AND DO NOT ANY EXPRESS OR IMPLIED, WARRANTIES OR CONDITIONS OF MERCHANTABILITY, FITNESS OR USAGE FOR PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.