SAP HANA and vPMEM

Table of Contents

1 Background....3
1.1 Prerequistes....3
1.2 Sizing....3
2 Configuring vPMEM....5
2.1 Configuring the LPAR profile for vPMEM....5
2.2 Managing the vPMEM volumes on the HMC....6
2.3 Preparing vPMEM volumes for use by SAP Hana....6
2.3.1 Direct Access for Files....6
2.4 Configure SAP Hana to use vPMEM....8
2.4.1 If a new installation....9
2.4.2 Automated rebuild and mount of vPMEM based file systems....9
2.5 Removing the vPMEM volumes....10
3 Appendix 1 - Appendix 1 – SAP Notes....12
4 References....13

Version details

Version	Date	Author	Description
1.0	15/09/20	Red	Initial design

1 Background

SAP HANA uses in-memory database technology that allows much faster access to data than was ever possible with hard disk technology on a conventional database. However one of the downsides is the amount of time required to load the data from disk into memory after a system restart. SAP has implemented some features to improve the load time, but it is still substantial for large databases.
System providers have also provided some solutions:
Intel has introduced Optane DC memory, where the data in memory is preserved when the system is powered down. Optane DC memory is positioned as a new tier of storage, lying between DRAM and Flash Storage in terms of performance.
IBM has introduced Virtual Persistent Memory (vPMEM), which is just the standard physical memory, but the Power Hypervisor creates a pool out of the already installed DRAM and keeps it persistent after the LPAR is shutdown. The data in this pool will be persistent as long as the Power Server itself is not powered down and has the added benefit that there is no performance loss or extra cost.

	VPMEM persistence	PMEM persistence
Application restart	Yes	Yes
LPAR restart	Yes	Yes
Physical server restart	No	Yes

1.1 Prerequistes

POWER9 processor-based systems with Firmware FW940 or later
Hardware Management Console (HMC) V9R1 M940 or later
System firmware FW940 or above
PowerVM level V3.1.1 or above
SUSE Linux 15 for SAP Applications 15 Service Pack 1 with

- kernel up to 4.12.14-197.26-default or later
- ndctl level 64.1-3.3.1

SAP HANA 2.0 SPS 03 Revision 35 (2.00.035)
SAP HANA 2.0 SPS 04 Revision 44 (adds new features for memory management)

1.2 Sizing

Before configuring vPMEM volumes for use with SAP HANA, a sizing exercise should be completed to determine the appropriate volume sizes necessary.
The vPMEM volumes should be as large as the anticipated main size of the column store plus spare capacity for growth and delta merge operations.

SAP note 2786237 details a number to tools to assist in proper sizing of persistent memory volumes. The SQL reports attached to the note can be used for an overview of the memory usage in the current system.

2 Configuring vPMEM

2.1 Configuring the LPAR profile for vPMEM

Each LPAR has an associated hardware page table (HPT) which resolves a LPAR effective addresses to real physical addresses in the hardware. The amount of memory that the HPT itself requires is based on the maximum DRAM that may be used by the partition and the HPT ratio. The HPT ratio is the ratio of the HPT size to the maximum memory value for the logical partition and can affect the performance of the logical partition. A small HPT might incur more CPU consumption as operating system might need to reload the entries in the HPT more frequently. The default HPT ratio for Linux LPARs is 1/128th of the maximum memory.
Currently, with Power Systems firmware FW940, HPT sizing is based on the LPAR setting for DRAM maximum memory. vPMEM memory usage is not included and any large vPMEM volumes will cause the HPT to be too small. It is recommended to use the hpt_for_vpmem.py script to set the LPAR profile parameters for appropriate HPT sizing.
The options for the script are as follows:

```
hpt_for_vpmem.py [options]*
-h, --help                    Displays the help text.
-m GB, --memory GB            Desired DRAM memory for the partition, in GB.
-l [Nx]GB, --lun [Nx]GB       VPMEM LUN size in GB, can be specified
                              multiple times. Can also be specified with a
                              preceding replication factor N, as in '3x1000'
                              for three 1000GB LUNs.
-i, --ibmi                    Partition type is IBMi.
-n, --linux                   Partition type is Linux.
-a, --aix                     Partition type is AIX
```

For example, asking for recommendations for an LPAR with 700GB desired memory and 4 vPMEM volumes of 1536GB

```
./hpt_for_vpmem.py --memory 700 --lun 4x1536 --linux
```

```
Inputs:
desired_memory_size     = 700GB
vpmem_size              = 6144GB
hpt_ratio               = 1/128 (7)
ppt_ratio               = 1/4096 (6)

Goals:
target_hpt_size = 64GB

Outputs:
max_memory_size   = 1281GB
hpt_ratio         = 1/32 (5)
ppt_ratio         = 1/1024 (4)
```

```
actual_hpt_size   = 64GB

ELMM Tree Structure:
elmm_base_address       = 4TB
PCI/VAS/XIVE            = 4TB..8TB
LUN 1536GB              = 8TB..10TB
LUN 1536GB              = 10TB..12TB
LUN 1536GB              = 12TB..14TB
LUN 1536GB              = 14TB..16TB
elmm_end_address        = 16TB

Recommendations:
Change the HPT ratio from 1/128 to 1/32.
Change the PPT ratio from 1/4096 to 1/1024.
Change the maximum memory size from 700GB to 1281GB.
```

The profile can then be changed with the chsyscfg command as follows:

```
chsyscfg -r prof -m <managed-system name> -i "name=<profile_name>,\
lpar_name=<partition_name>,max_mem=1281,hpt_ratio=1:32,ppt_ratio=1:1024”
```

2.2 Managing the vPMEM volumes on the HMC

vPMEM volumes are configured at the LPAR level and the LPAR cannot be in an activated state.

From the HMC Menu, Select the Partition, in the Properties menu on the LHS, select “Persistent Memory” and ‘Add”
Specify the name for the persistent memory volume and the size (in MB) – should be a multiple of the LMB size (typically 256MB). For SAP Hana, check the affinity box. Select “OK”

2.3 Preparing vPMEM volumes for use by SAP Hana

By default SAP HANA places partition data into files in an XFS Filesystem mounted with the DAX option. Now we build this filesystem on the vPMEM volumes. After defining vPMEM volumes on the HMC and booting the LPAR, the volumes are presented as non-volatile DIMM devices,
/dev/nmem<#>, by the OS.

2.3.1 Direct Access for Files

For block devices that are memory-like, the page cache pages would be unnecessary copies of the original storage. The DAX code removes the extra copy by performing reads and writes directly to the storage device. For file mappings, the storage device is mapped directly into userspace.

The ndctl tool is used to list the vPMEM volumes

```
ndctl list --dimms
```

```
[
      {
            "dev":"nmem1"
      },
      {
            "dev":"nmem0"
      }
]
```

Can also list the associated blocks

```
ndctl list --bus all
[
      {
            "provider":"ibm,persistent-memory:ibm,pmemory@44108001",
            "dev":"ndbus1"
      },
      {
            "provider":"ibm,persistent-memory:ibm,pmemory@44104001",
            "dev":"ndbus0"
      }
]
```

To prepare the volumes use ndctl tool to create the required pmem regions and namespaces on these devices.
For example for 2 volumes (/dev/nmem0 and /dev/nmem1) do

```
for i in 0 1
do
      ndctl disable-region region$i
      ndctl zero-labels nmem$i
      ndctl init-labels nmem$i
      ndctl enable-region region$i
      ndctl create-namespace -r region$i
done
```

Now the memory block devices have been prepared, and seen by the OS as /dev/pmen#. Can be listed by ndctl

```
ndctl list --namespaces --type=pmem
[
      {
            "dev":"namespace1.0",
            "mode":"fsdax",
            "map":"dev",
            "size":402787401728,
            "uuid":"68080d19-6890-4bce-b612-6b4af4e164b1",
            "sector_size":512,
            "align":16777216,
            "blockdev":"pmem1"
      },
```

```
      {
            "dev":"namespace0.0",
            "mode":"fsdax",
            "map":"dev",
            "size":401713659904,
            "uuid":"57d3aea8-b8cd-410c-960e-047b8cc03949",
            "sector_size":512,
            "align":16777216,
            "blockdev":"pmem0"
      }
]
```

Now create an XFS File system with the DAX option for each pmem namespace. As DAX option skips the page cache and uses the file system blocks directly, we need to have the block size the same as the PAGESIZE (64K on Power)

```
for i in 0 1
do
      mkdir -p /hana/shared/pmem /pmem$i
      mkfs.xfs /dev/pmem$i -b size=64k -s size=512
      mount -o dax /dev/pmem$i /hana/shared/pmem /pmem$i
done
chown -R <sid>adm:sapsys /data/hanapm
chmod -R 0600 /data/hanapm
```

Note: pmem file system mount points cannot be created under other file system mount points (other than /) as SAP HANA will not use the DAX attribute correctly.

Note: Note also the blockdev name, e.g. pmem0 and pmem1 from above, may change on reboot. For any automated mounting of the associated filesystems, it is recommended to use the filesystem UUID (*blkid*)

2.4 Configure SAP Hana to use vPMEM

SAP HANA configuration files stored on the server at the following locations according to layer:

Default:
/usr/sap/<SID>/HDB<instance>/exe/config (read only)
System:
<sapmnt>/<SID>/SYS/global/hdb/custom/config
Database:
<sapmnt>/<SID>/SYS/global/hdb/custom/config/DB_<dbname>
Host:
/usr/sap/<SID>/HDB<instance>/<hostname>

By default, SAP HANA is defined at the host level to use persistent memory volumes. All HANA services managed by a single SAP HANA Global Allocation Limit (GAL) will share a set of persistent memory volumes.

Use the basepath_persistent_memory_volumes parameter to specify the pmem filesystem in HANA global.ini configuration file, e.g.:

```
...
[persistence]
basepath_datavolumes = /hana/shared/data/JE6
basepath_logvolumes = /hana/shared/log/JE6
basepath_persistent_memory_volumes = \
      /hana/shared/pmem/pmem0/JE6;/hana/shared/pmem/pmem1/JE6;
```

Activate persistent memory storage for database in HANA indexserver.ini configuration file.

```
...
[persistent_memory]
table_default = ON
```

Note: that this setting may be overridden by the preference settings on the table, partition or column level.

2.4.1 If a new installation

For new installations, add the following parameters to hdblcm

```
--use_pmem --pmempath=<path to pmemX>[:<path to pmemY>...]
```

To confirm that vPMEM is being used
The following query can be used to verify that the vPMEM based file system is being used as expected:

```
hdbsql> select * from M_PERSISTENT_MEMORY_VOLUMES where PORT=3<instance #>03
```

for example:

```
hdbsql> select * from M_PERSISTENT_MEMORY_VOLUMES where PORT=31003
HOST,PORT,VOLUME_ID,NUMA_NODE_INDEX,PATH,FILESYSTEM_TYPE,IS_DIRECT_ACCESS_SUPPORTED,TOTAL_SIZE,USED_SIZE
"lsh30117",31003,3,0,"/hana/shared/pmem/pmem0/JE6/mnt00001/hdb00003.00003","xfs","true",401517510656,15582494720
"lsh30117",31003,3,1,"/hana/shared/onen/pmem1/JE6/mnt00001/hdb00003.00003","xfs","true",402590728192,15930228736
```

This shows that HANA has found and is using 2 persistent memory-based XFS filesystems. One filesystem is backed by memory on NUMA node 0, while the other is backed by memory on NUMA node 1.

2.4.2 Automated rebuild and mount of vPMEM based file systems

When maintaining a HANA system, activities such as restarting the operating system (e.g. for applying security fixes) or restarting the managed system are on occasion required. When using vPMEM with HANA, some additional steps must be taken before restarting HANA. In case of an OS reboot, the earlier created file systems must be remounted. In case of a managed system restart with Power Off the underlying files systems also need to be rebuilt.
To simplify and automate those actions, a convenient startup script is available. The vpmem_hana_startup.sh script assists and automates the process of verifying the vPMEM based file systems, mounting the file systems and updating the HANA configuration file.

Create a configuration input file vpmem_hana.cfg

```
[
      {
      "sid" : "<HANA instance name>"
      ,"puuid": "<parent vpmem volume uuid>"
      ,"mnt" : "<filesystem path to mount vpmem filesystems under>"
      }
]
```

Confirm the UUID of the pmem block devices with the script.
It will:

- scan the configuration file to determine the parent UUID of the vPMEM volumes.
- Search the device tree to locate the vPMEM devices associated with the UUID.
- For each child volume, check whether valid filesystems exist
- If no valid file systems found, format them with an XFS filesystem.
- Mount each of the filesystems under a mount point representing their NUMA associativity.
- Update the HANA configuration file to reflect where the vPMEM devices are mounted for each NUMA domain.

```
./vpmem_hana_startup.sh -p
/sys/devices/ndbus0/region0/of_node/ibm,unit-parent-guid \
      "71043c70-3d8f-42fa-8d7d-2828c04666f5"
/sys/devices/ndbus1/region1/of_node/ibm,unit-parent-guid \
      "71043c70-3d8f-42fa-8d7d-2828c04666f5"
```

Use the systemd service to start the script on OS start
Put the script and the configuration file in /usr/sap/vpmem
create the /etc/systemd/system/vpmem_hana.service as:

```
[Unit]
Description=Virtual PMEM SAP HANA Startup Script

[Service]
Type=oneshot
ExecStart=/bin/sh -c "/usr/sap/vpmem/vpmem_hana_startup.sh"
```

```
[Install]
WantedBy=multi-user.target
```

Start the service now and on reboot

```
systemctl start vpmem_hana.service
systemctl enable vpmem_hana.service
```

2.5 Removing the vPMEM volumes

Note: You can use the *ndctl destroy-namespaces* command to remote the pmem volumes.
For example

```
for i in 0 1
do
        ndctl destroy-namespace namespace${i}.0 -f
done
```

3 Appendix 1 - Appendix 1 – SAP Notes

SAP note	Title
2618154	SAP HANA Persistent Memory – Release Information
2700084	FAQ:SAP HANA Persistent Memory
2786237	Sizing SAP HANA with Persistent Memory
2175606	HANA: How to set allocation limit for tenant databases

4 References

SAP HANA Administration Guide – Administration Guide: Persistent Memory
(https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.04/en-US/1f61b13e096d4ef98e62c676debf117e.html)

IBM product documentation:
Managing persistent memory volume
https://www.ibm.com/support/knowledgecenter/en/9040-MR9/p9efd/p9efd_lpar_pmem_settings.htm

SAP product documentation:
Persistent Memory - SAP HANA Administration Guide for SAP HANA Platform
https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.04/en-US/1f61b13e096d4ef98e62c676debf117e.html

SAP HANA and PowerVM Virtual Persistent Memory
Planning and Implementation Guide
Jim Nugen Olaf Rutz

Using IBM POWER9 PowerVM Virtual Persistent Memory for SAP HANA with SUSE Linux - Posted on January 9, 2020 (Jay Kruemcke)
https://kruemcke.wordpress.com/2020/01/09/using-ibm-power9-powervm-virtual-persistent-memory-for-sap-hana-with-suse-linux/